

DS3500

Advanced Programming with Data

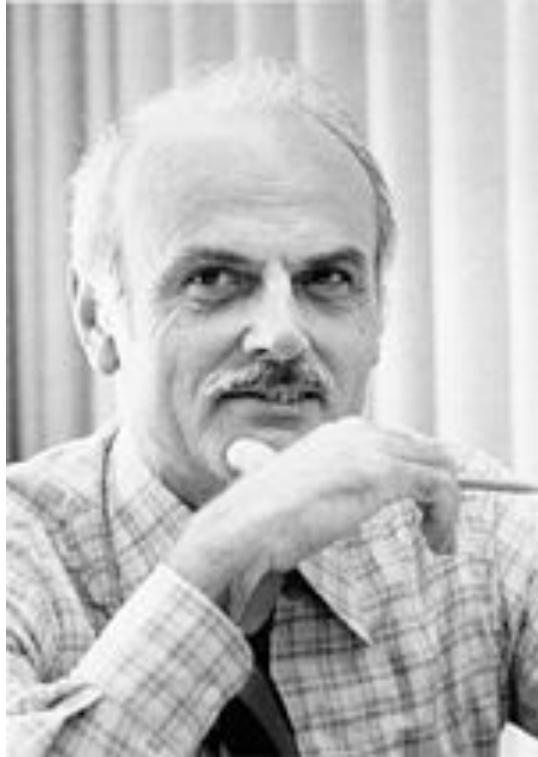
John Rachlin
Northeastern University



Northeastern University

Database Programming





Edgar Frank Codd
1923-2003

Information Retrieval

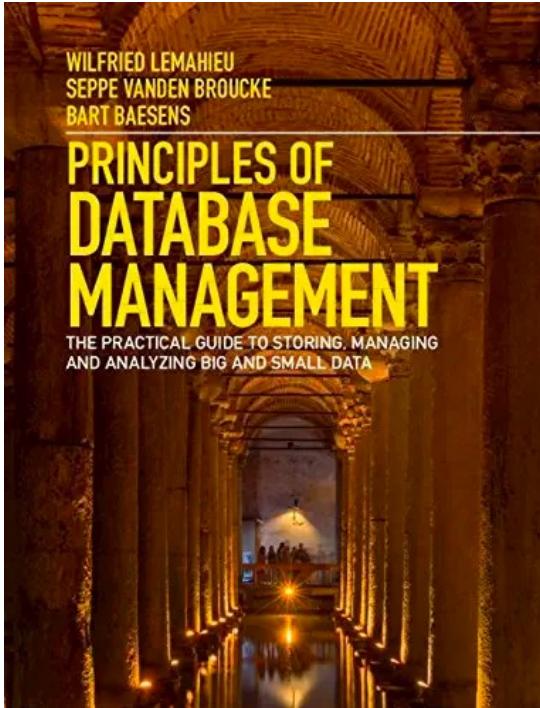
A Relational Model of Data for Large Shared Data Banks

E. F. CODD

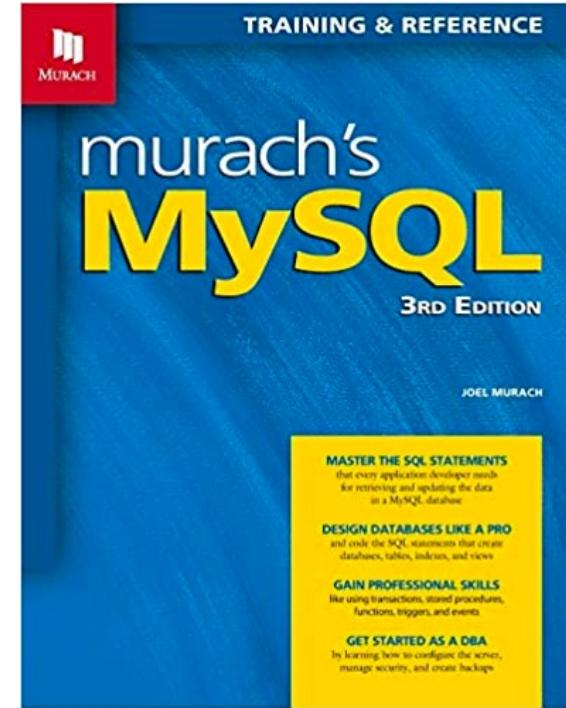
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

Textbooks



Principles of Database Management (1ed)
by LeMahieu, vanden Broucke and Baesens
Recommended



Murach's MySQL by Joel Murach (3rd edition)
(2nd ed OK)
Highly recommended



What is a Database?

- Shared collection of **logically related** data (and a description of this data), designed to meet the information needs of an organization.
- System **catalog** (metadata) provides description of data to enable program–data independence.
- Logically related data comprises entities, attributes, and relationships of an organization’s information.
- In a relational database, the data is represented as a **collection of tables**.

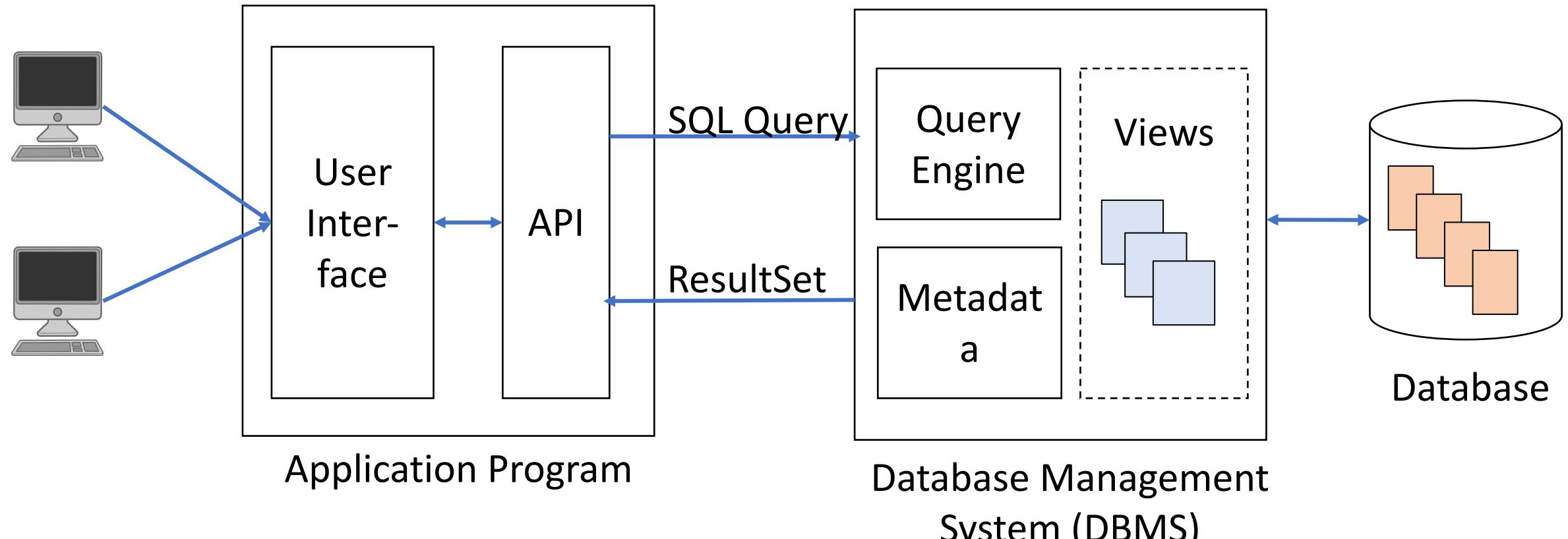


Database Management System (DBMS)

- A software system that enables users to define, create, maintain, and control access to the database.
- (Database) application program: a computer program that interacts with database by issuing an appropriate request (SQL statement) to the DBMS.
- A *Relational* Database Management System (RDBMS or RDB) uses the relational model: i.e., it uses tables to represent entities. This remains the dominant model in industry today, but other approaches have emerged.



Database Application



A Simple Relational Data Model

Name	Salary	Branch
John White	30000	22 Deer Rd, London
Ann Beech	12000	163 Main St, Glasgow
David Ford	18000	163 Main St, Glasgow
Mary Howe	9000	16 Argyll St, Aberdeen
Susan Brand	24000	163 Main St. Glasgow
Julie Lee	9000	22 Deer Rd, London

Discussion: What are the disadvantages of this design?



A Better Relational Data Model

Branch

branchNo	street	city	postCode
B005	22 Deer Rd	London	SW1 4EH
B007	16 Argyll St	Aberdeen	AB2 3SU
B003	163 Main St	Glasgow	G11 9QX
B004	32 Manse Rd	Bristol	BS99 1NZ
B002	56 Clover Dr	London	NW10 6EU

Staff

staffNo	fName	IName	position	sex	DOB	salary	branchNo
SL21	John	White	Manager	M	1-Oct-45	30000	B005
SG37	Ann	Beech	Assistant	F	10-Nov-60	12000	B003
SG14	David	Ford	Supervisor	M	24-Mar-58	18000	B003
SA9	Mary	Howe	Assistant	F	19-Feb-70	9000	B007
SG5	Susan	Brand	Manager	F	3-Jun-40	24000	B003
SL41	Julie	Lee	Assistant	F	13-Jun-65	9000	B005



Relational Data Model

Entities

branchNo	street	city	postCode
B005	22 Deer Rd	London	SW1 4EH
B007	16 Argyll St	Aberdeen	AB2 3SU
B003	163 Main St	Glasgow	G11 9QX
B004	32 Manse Rd	Bristol	BS99 1NZ
B002	56 Clover Dr	London	NW10 6EU

Staff

staffNo	fName	IName	position	sex	DOB	salary	branchNo
SL21	John	White	Manager	M	1-Oct-45	30000	B005
SG37	Ann	Beech	Assistant	F	10-Nov-60	12000	B003
SG14	David	Ford	Supervisor	M	24-Mar-58	18000	B003
SA9	Mary	Howe	Assistant	F	19-Feb-70	9000	B007
SG5	Susan	Brand	Manager	F	3-Jun-40	24000	B003
SL41	Julie	Lee	Assistant	F	13-Jun-65	9000	B005

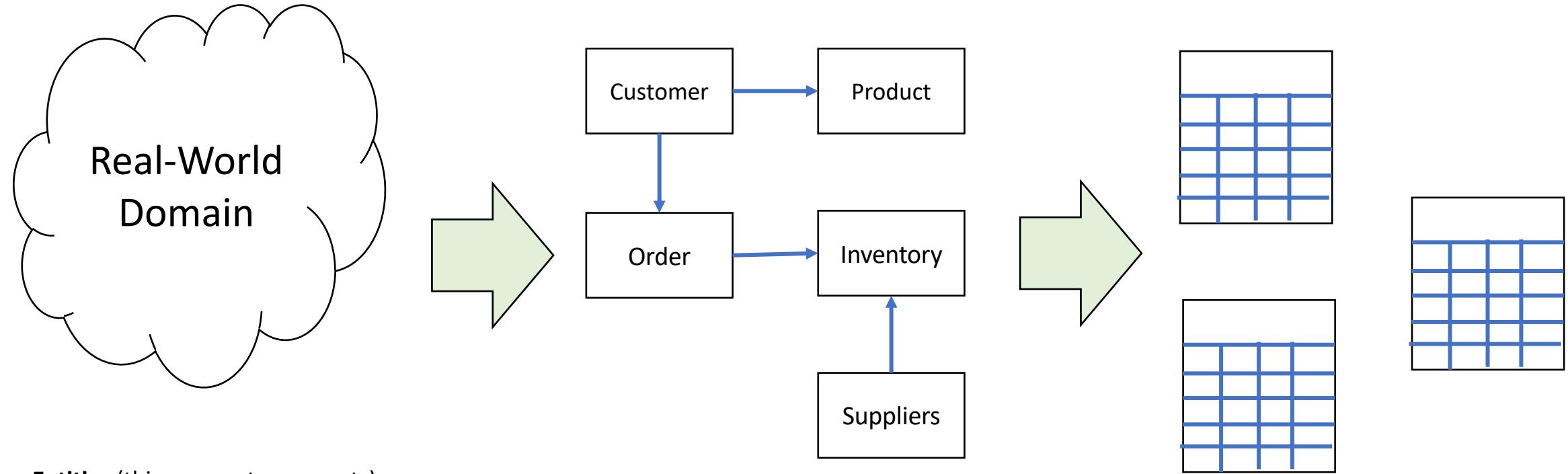
A column or **attribute** or **field** providing more details about the particular entity.

- Name
- Datatype
- Other attributes

A row or **record** describing one instance of the entity



The Modeling Process

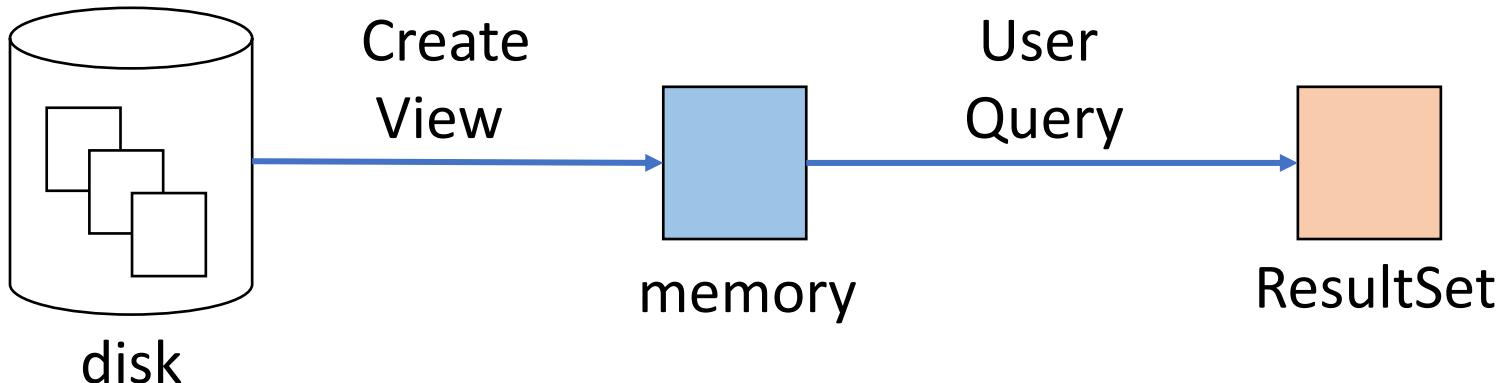


- **Entities** (things, events, concepts)
- **Attributes** (annotations on entities)
- **Relationships** (connections or associations between entities)

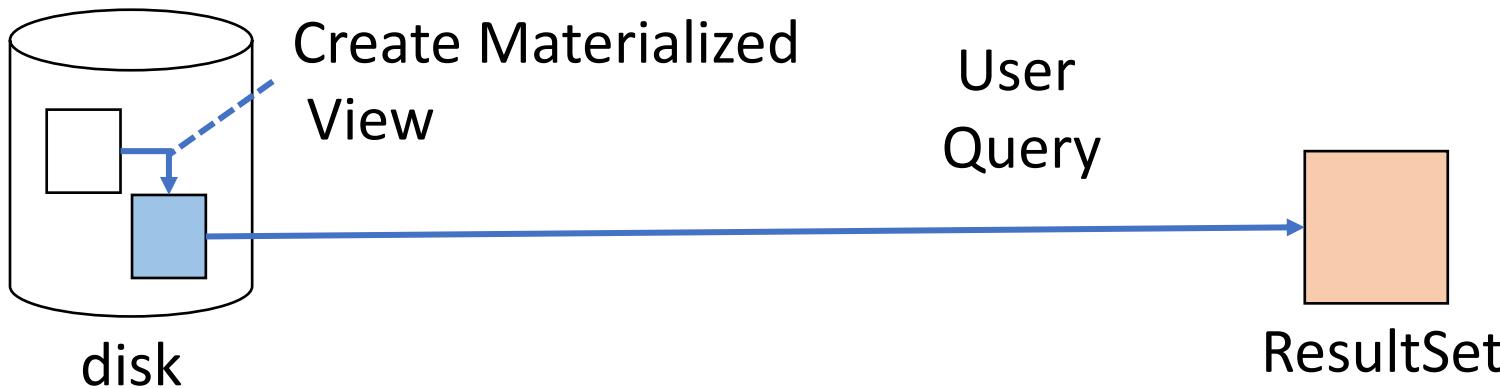


How Views Work

Standard View (MySQL)



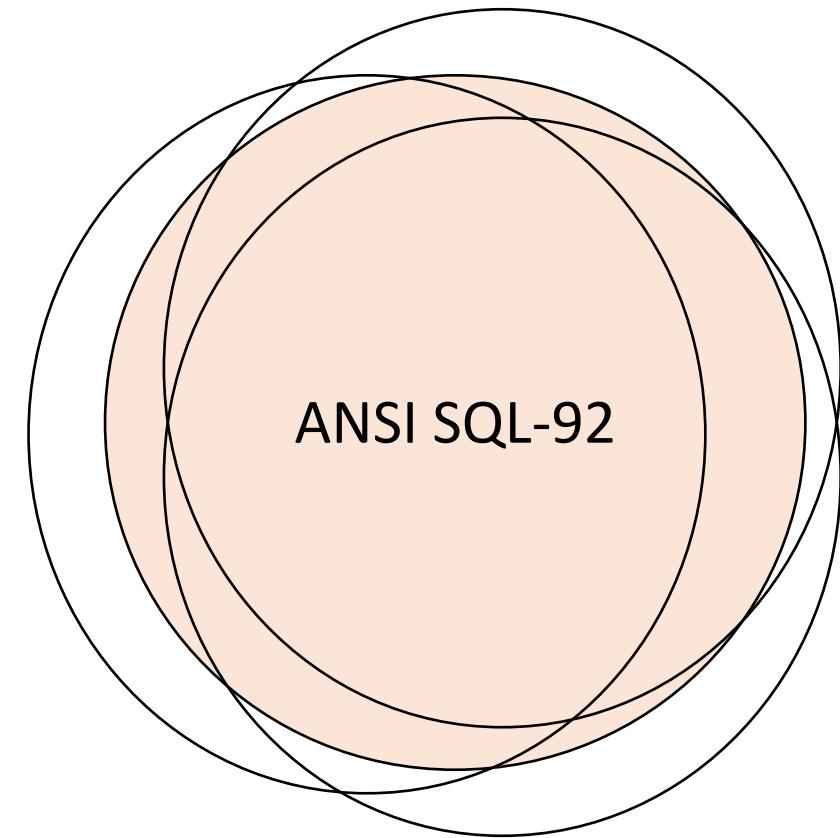
Materialized View (Some DBMSs)



- View is a *virtual* table
- Always up-to-date
- Every access regenerates the create-view operation in the background
- View is a *physical* table on disk
- View is not always up-to-date: Needs to be periodically synced.
- Subsequent user queries are faster, however because creation is done only once.

SQL: The language of databases

- SQL = Structured Query Language
- Pronounced *either* S.Q.L. or *See*-quel (both are perfectly acceptable.)
- The language was *standardized* over time (SQL-86, SQL-89, SQL-92) but most vendors do not implement the standard completely or they introduce vendor-specific features / dialects.



One standard, many dialects



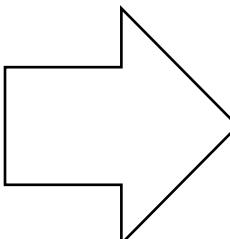
What is *structured* about SQL?

Simplified syntax for selecting data:

```
SELECT <Columns to retrieve>
FROM   <Source tables>
WHERE  <Conditions on individual rows>
ORDER BY <Sorting criteria>
```

An example of how SQL is a declarative language

```
SELECT invoice_id, invoice_total
FROM   invoices
WHERE  invoice_total > 100.0
ORDER BY invoice_total
```



	invoice_id	invoice_total
▶	1	125.00
	4	2199.99



Procedural Languages for DB are numerous

Source	Common name	Full name
ANSI/ISO Standard	SQL/PSM	SQL/Persistent Stored Modules
Interbase / Firebird	PSQL	Procedural SQL
IBM DB2	SQL PL	SQL Procedural Language (implements SQL/PSM)
IBM Informix	SPL	Stored Procedural Language
IBM Netezza	NZPLSQL ^[18]	(based on Postgres PL/pgSQL)
Microsoft / Sybase	T-SQL	Transact-SQL
Mimer SQL	SQL/PSM	SQL/Persistent Stored Module (implements SQL/PSM)
MySQL	SQL/PSM	SQL/Persistent Stored Module (implements SQL/PSM)
MonetDB	SQL/PSM	SQL/Persistent Stored Module (implements SQL/PSM)
NuoDB	SSP	Starkey Stored Procedures
Oracle	PL/SQL	Procedural Language/SQL (based on Ada)
PostgreSQL	PL/pgSQL	Procedural Language/PostgreSQL Structured Query Language (implements SQL/PSM)
Sybase	Watcom-SQL	SQL Anywhere Watcom-SQL Dialect
Teradata	SPL	Stored Procedural Language
SAP	SAP HANA	SQL Script

Source:
<https://en.wikipedia.org/wiki/SQL>



Most common datatypes

- CHAR(n), VARCHAR(n)
- INT (SIGNED or UNSIGNED)
- FLOAT, DOUBLE, DECIMAL(m,d)
- DATE, TIME, DATETIME
- BLOB, TEXT (CLOB)
- ENUM



Creating Table Examples

A very basic table with just four columns.

```
USE demo;
```

```
DROP TABLE IF EXISTS users;
```

```
CREATE TABLE users (
    user_id INT,
    username VARCHAR(20),
    dob DATE,
    class ENUM ('Basic', 'Premium')
);
```



What is a PRIMARY KEY?

- A **PRIMARY KEY** is a value or set of values that uniquely identifies a single row in your database table.
- Primary keys cannot be NULL (empty / unassigned)
- Primary keys must be unique
- Primary keys are usually INTs but they could be other datatypes



A Relational Data Model

Branch

branchNo	street	city	postCode
B005	22 Deer Rd	London	SW1 4EH
B007	16 Argyll St	Aberdeen	AB2 3SU
B003	163 Main St	Glasgow	G11 9QX
B004	32 Manse Rd	Bristol	BS99 1NZ
B002	56 Clover Dr	London	NW10 6EU

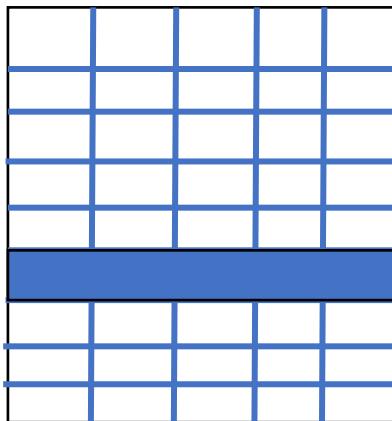
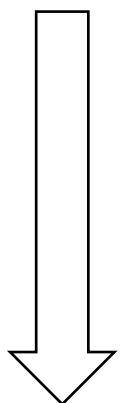
Staff

staffNo	fName	IName	position	sex	DOB	salary	branchNo
SL21	John	White	Manager	M	1-Oct-45	30000	B005
SG37	Ann	Beech	Assistant	F	10-Nov-60	12000	B003
SG14	David	Ford	Supervisor	M	24-Mar-58	18000	B003
SA9	Mary	Howe	Assistant	F	19-Feb-70	9000	B007
SG5	Susan	Brand	Manager	F	3-Jun-40	24000	B003
SL41	Julie	Lee	Assistant	F	13-Jun-65	9000	B005



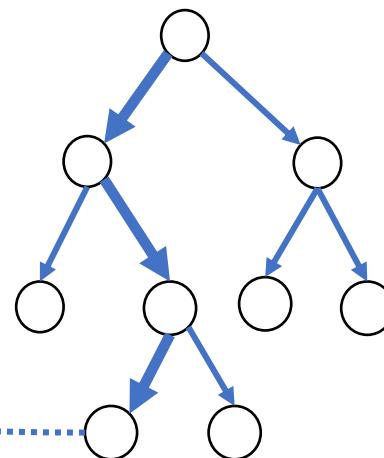
PRIMARY KEY: What's it good for?

The database management system builds an **index** for the table around the primary key so that, given the value of the primary key, it can look up the row containing the rest of the values very quickly.



$O(n)$

Without a primary Key
we must scan row-by-row
looking for the right value



$O(\log n)$

With Primary Key
we look up the the row location
by searching through a tree.



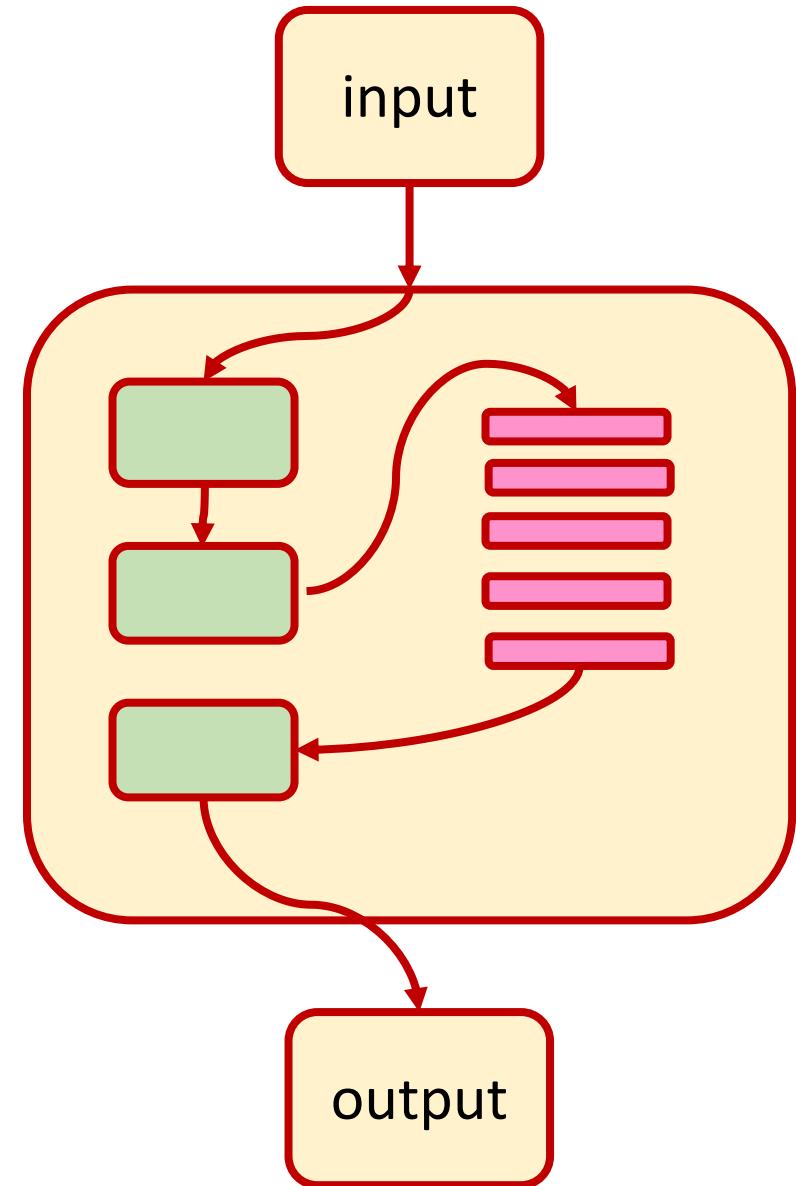
Object-Oriented Programming (OOP)



The procedural paradigm

Programs are *recipes*: a series of statements that transform our data into visualizations and insight.

In procedural programming, we manage complexity by being **modular**: adhering to top-down programming practices that break down difficult tasks into a sequence of more manageable sub-tasks.



Recipes are a powerful metaphor for data science!



start with
raw data



load
data



data refinement
(munging)



build data
structures

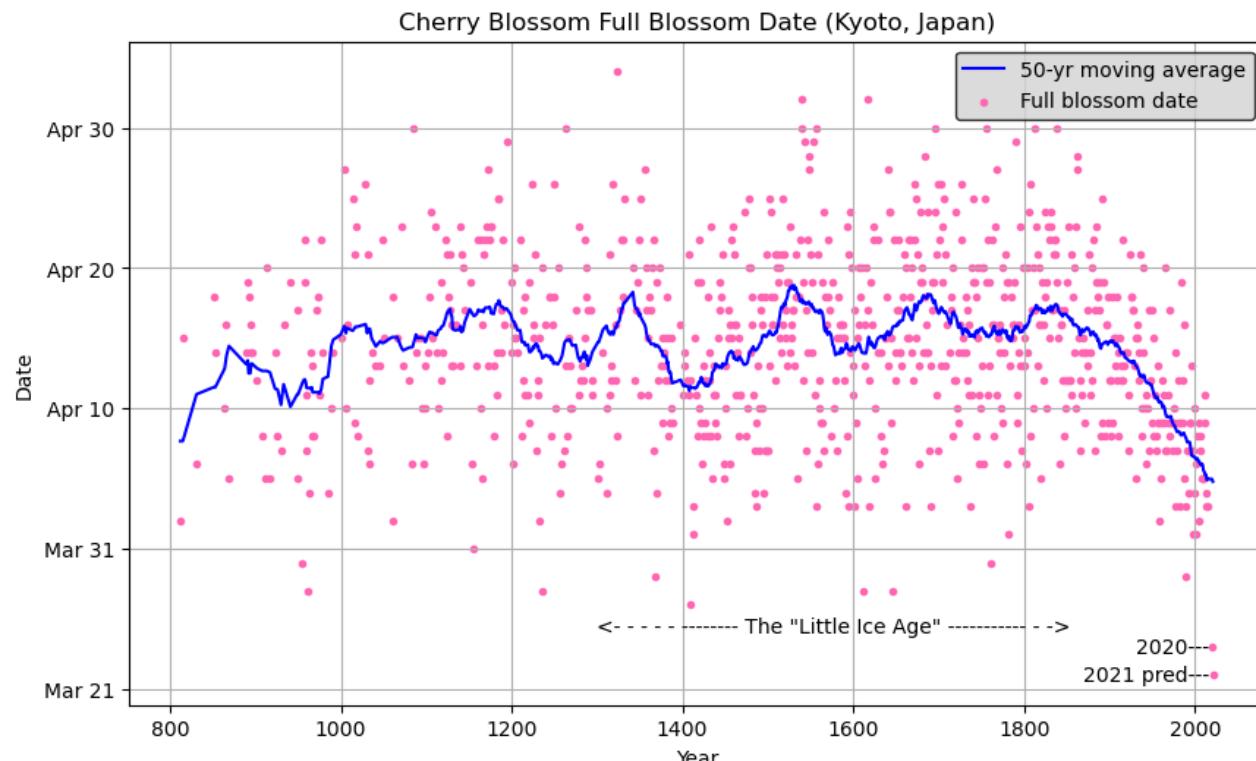
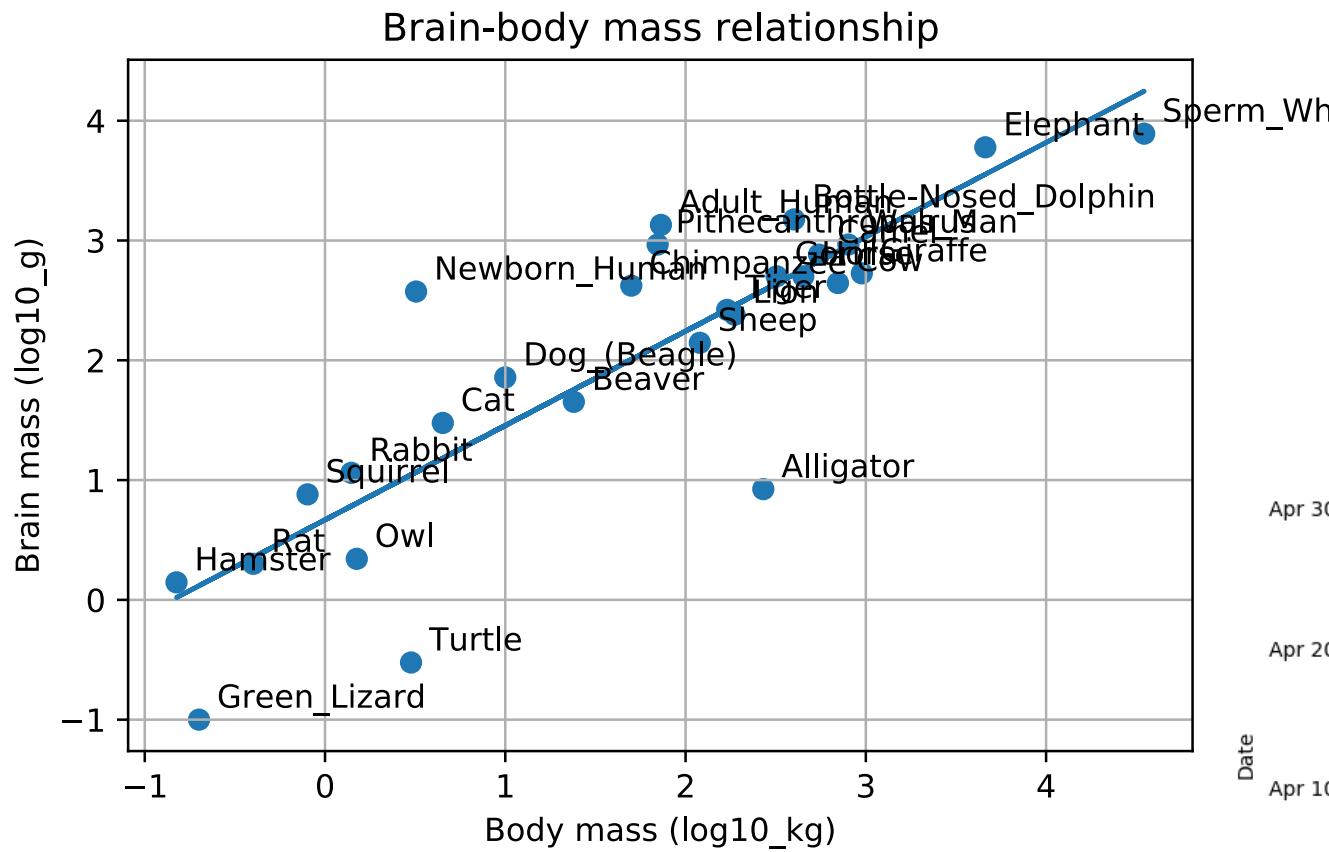


processing
and
visualization

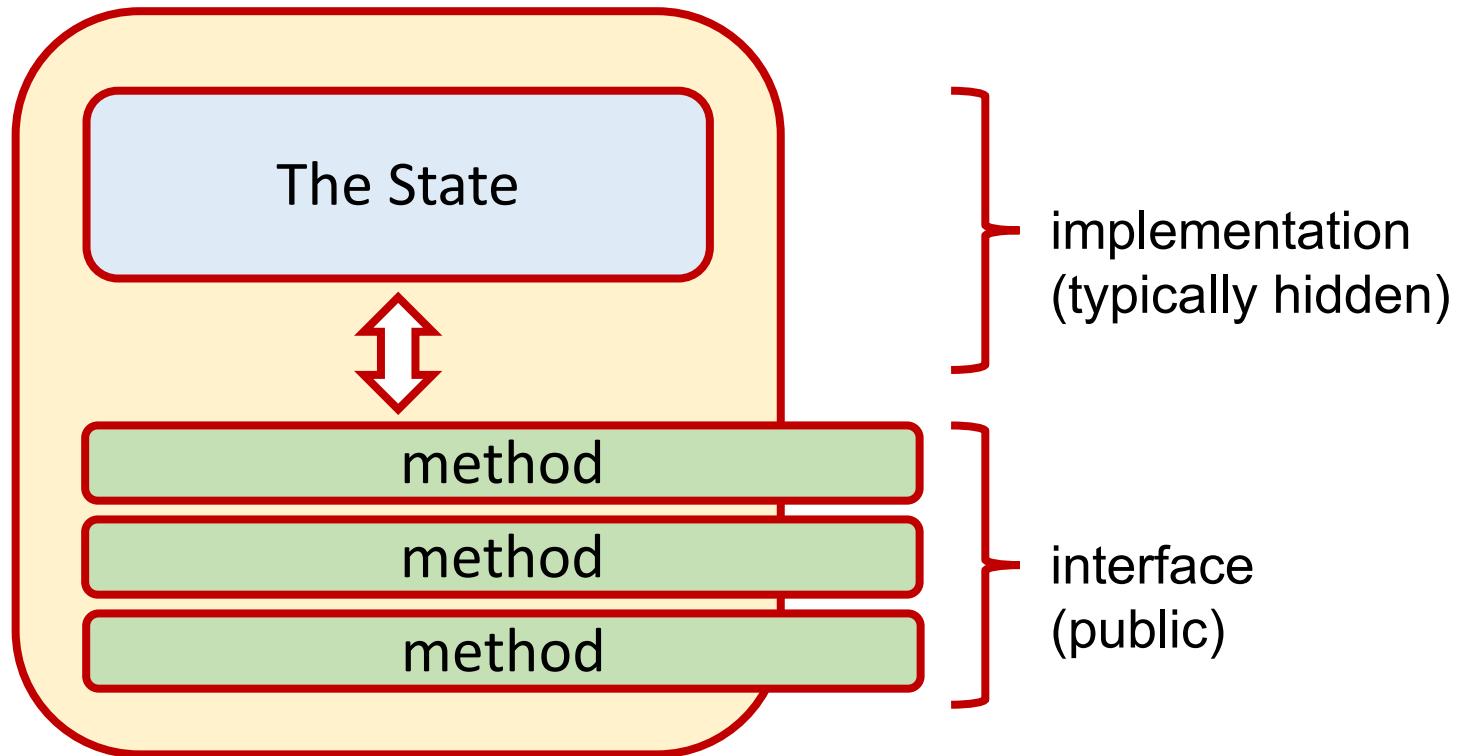


don't forget
to label your
axes!

...and it's adequate for a wide range of use-cases!



The object paradigm (more later!!)



Objects have fields and attributes that constitute the **state** of the object.

The state is accessed and modified through various **methods** that constitute the object's **interface**.

The layout and organization of the state (i.e., the **implementation**) is usually shielded from the user.

Beautiful Code (and also the DS2500 HW grading rubric!)

- **Modular:** The program is broken down into small reusable functions that are more easily implemented, tested, and verified.
- **Reusable:** Functions are parameterized for maximum flexibility. Functions that go together are packaged together in a file. You aren't re-inventing the wheel.
- **Readable:** Your code is well-documented so that another programmer can understand your thought process. (Your future self will thank you!) Every function or method is documented including a specification of expected inputs and outputs. Don't go overboard!
`x = 5 # assign 5 to x - not a useful comment!!`
- **Efficient:** Uses data-structures appropriately. Implements algorithms that meet performance requirements without being unnecessarily complex.
- **Correct:** Satisfies the requirements of the project or assignment.



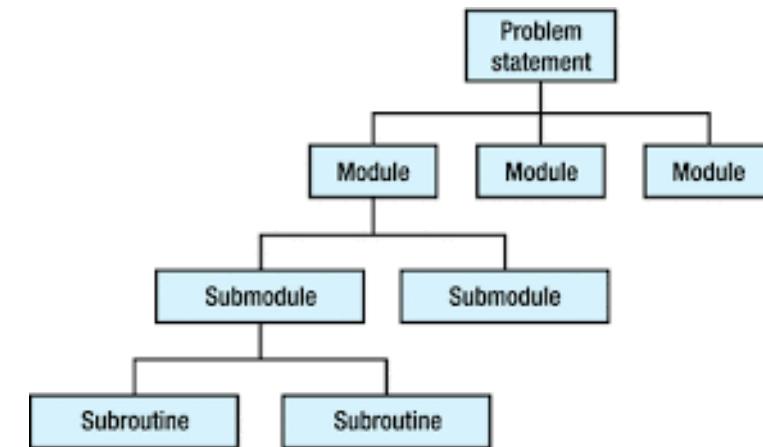
Starting with modular design

- Modular programming is a [software design](#) technique that emphasizes separating the functionality of a [program](#) into independent, interchangeable **modules**, such that each contains everything necessary to execute only one aspect of the desired functionality.

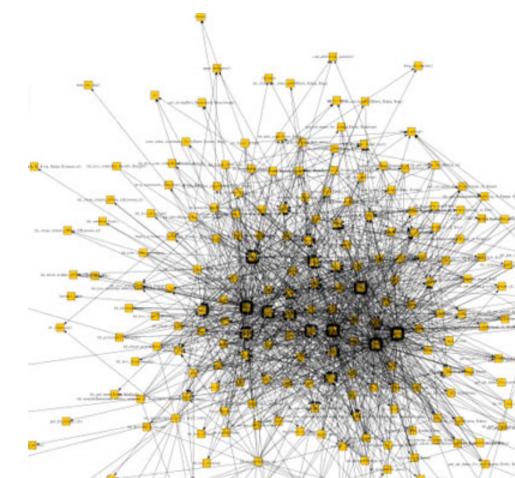
(Wikipedia)

- Modules are usually small (<100 lines of code)
- Modules are *re-useable* (if well-written)
- Modular programming limits cross-dependencies

modular code



Spaghetti code



Regular (Traditional) Packages in Python

- Regular packages are traditional packages as they existed in Python 3.2 and earlier.
- A regular package is typically implemented as a directory containing an `__init__.py` file. When a regular package is imported, this `__init__.py` file is implicitly executed, and the objects it defines are bound to names in the package's namespace.
- The `__init__.py` file can contain the same Python code that any other module can contain, and Python will add some additional attributes to the module when it is imported.

```
parent/
    __init__.py
    one/
        __init__.py
    two/
        __init__.py
    three/
        __init__.py
```

Importing `parent.one` will implicitly execute `parent/__init__.py` and `parent/one/__init__.py`. Subsequent imports of `parent.two` or `parent.three` will execute `parent/two/__init__.py` and `parent/three/__init__.py` respectively.



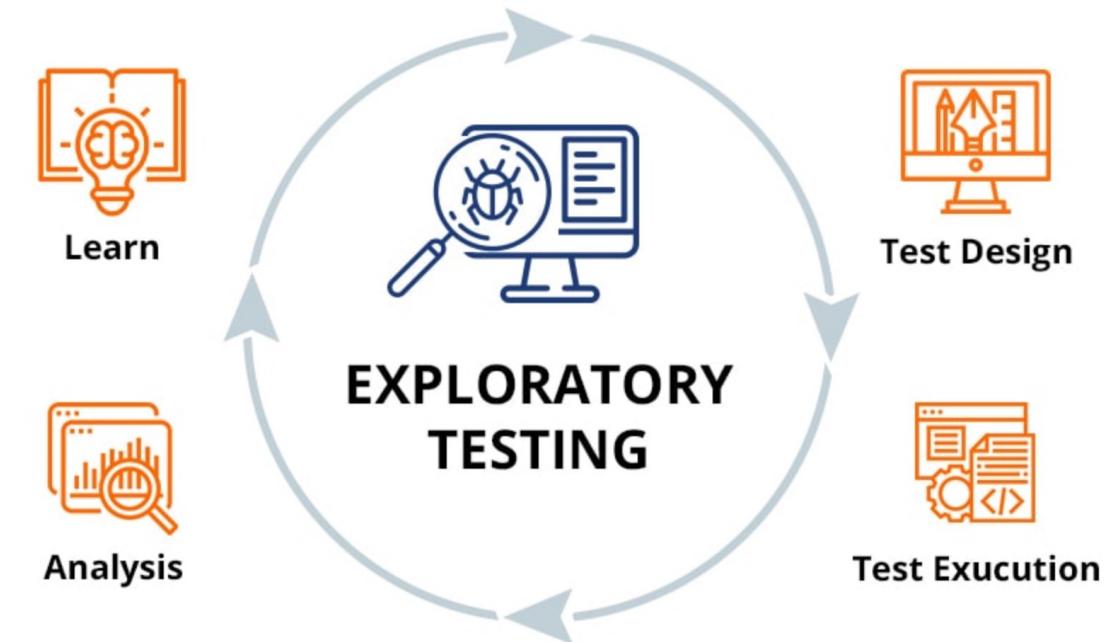
Testing

Exploratory Testing vs. Test-Driven Development



Exploratory (Manual) Testing

- Running a program against an unstructured (possibly incomplete) set of cases in order to validate a list of features is known as *exploratory* testing. This is *testing without a plan*.
- Every change requires manual re-testing to make sure nothing has broken.
- Dependent on the skill of the tester to ensure complete coverage.



Automated Testing

- Automated Testing involves the execution of a *test plan*.
- Python has built-in tools to support this.
- A *test runner* executes a series of pre-defined tests and compares results against expectations.
- **unit test:** Testing a single component, function, or module.
- **integration test:** Testing multiple components and their interactions.



Exploratory vs. Automated Testing

	EXPLORATORY TESTING	AUTOMATED TESTING
Documentation	not much need of documentation	documentation is essential
Lead-in time	little or no lead-in time	lead-in time is required
Test scripts	not much efforts for test script creation	test scripts are important
Test coverage	no measurement	can be measured
Based on	tester's understanding of a product	specifications
Reproducibility	low	high



pytest: A framework for python testing (pytest.org)



pytest

Table Of Contents

[Home](#)[Install](#)

Installation and Getting Started

Pythons: Python 3.6, 3.7, 3.8, 3.9, PyPy3

Platforms: Linux and Windows

PyPI package name: [pytest](#)

Documentation as PDF: [download latest](#)

pytest is a framework that makes building simple and scalable tests easy. Tests are expressive and readable—no boilerplate code required. Get started in minutes with a small unit test or complex functional test for your application or library.

```
$ conda install pytest
```

```
$ conda install pytest-cov
```

```
$ conda list | grep pytest
```



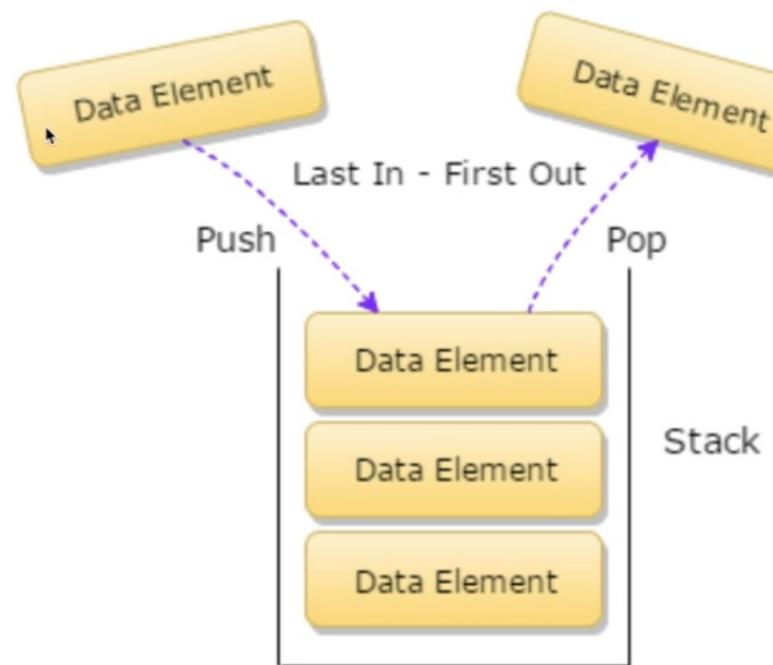
Northeastern University

Stacks and Queues

Stack (First In, Last out)

push: Add an element to the top of the stack

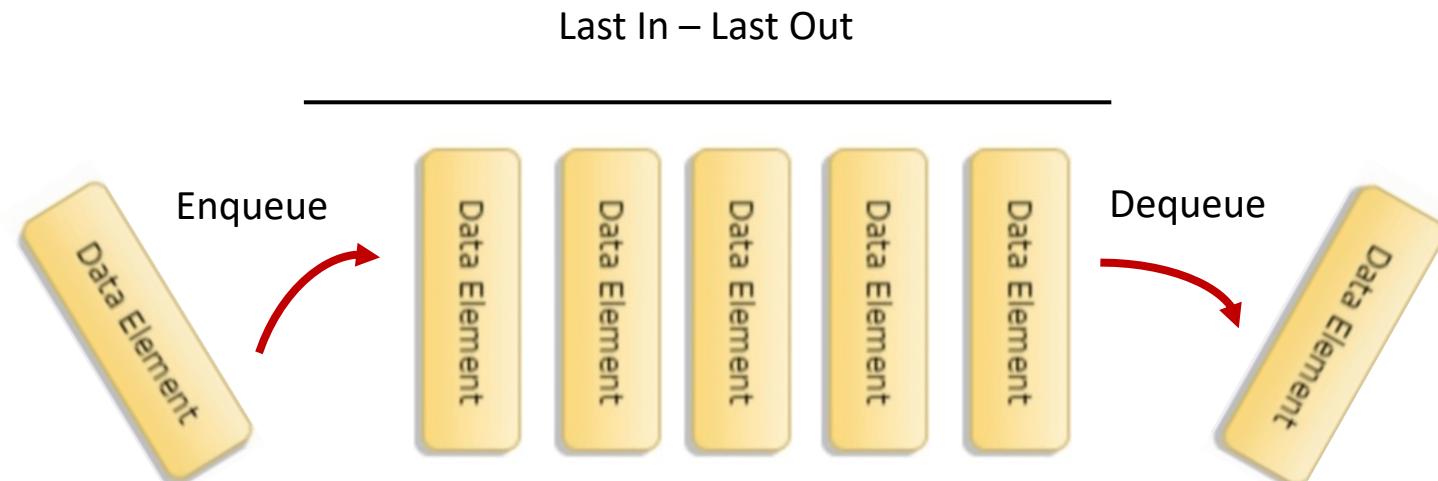
pop: Remove an element from the top of the stack



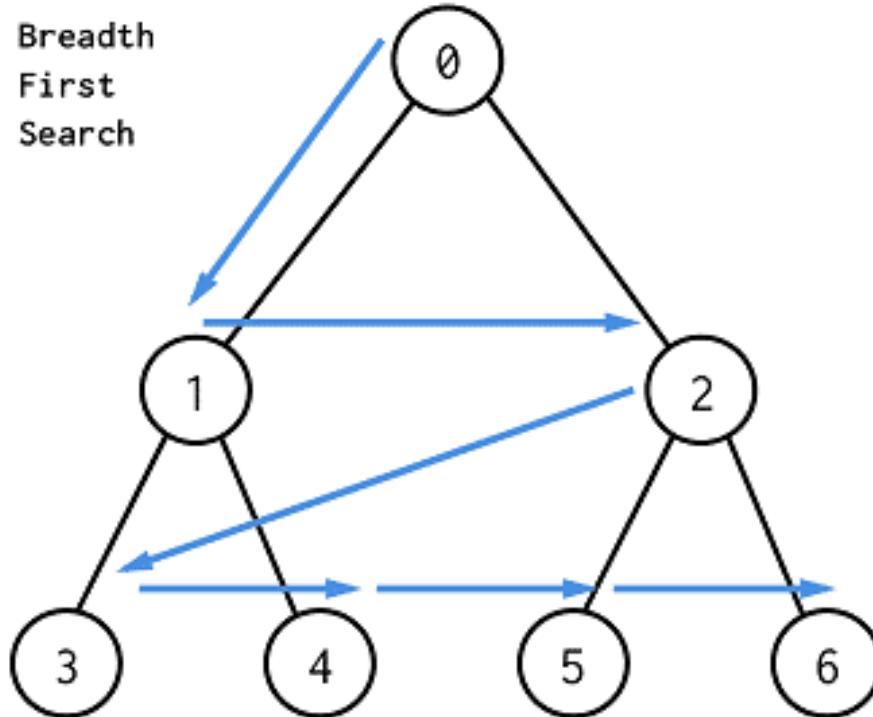
Queue (First In, First out)

enqueue: Add an element to the front of the queue

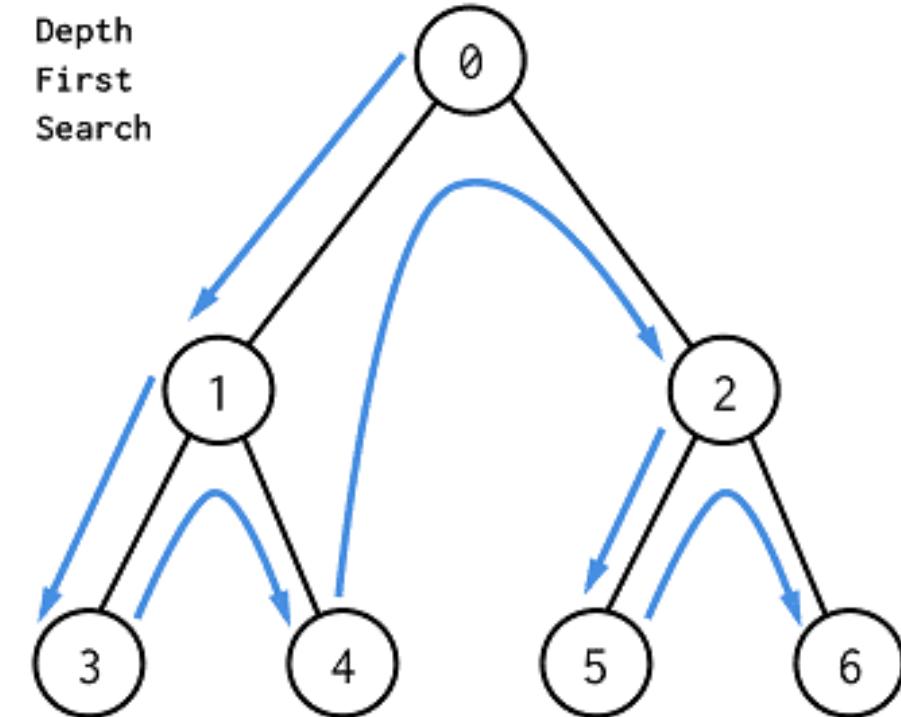
dequeue: Remove an element from the end of the queue



BFS vs. DFS



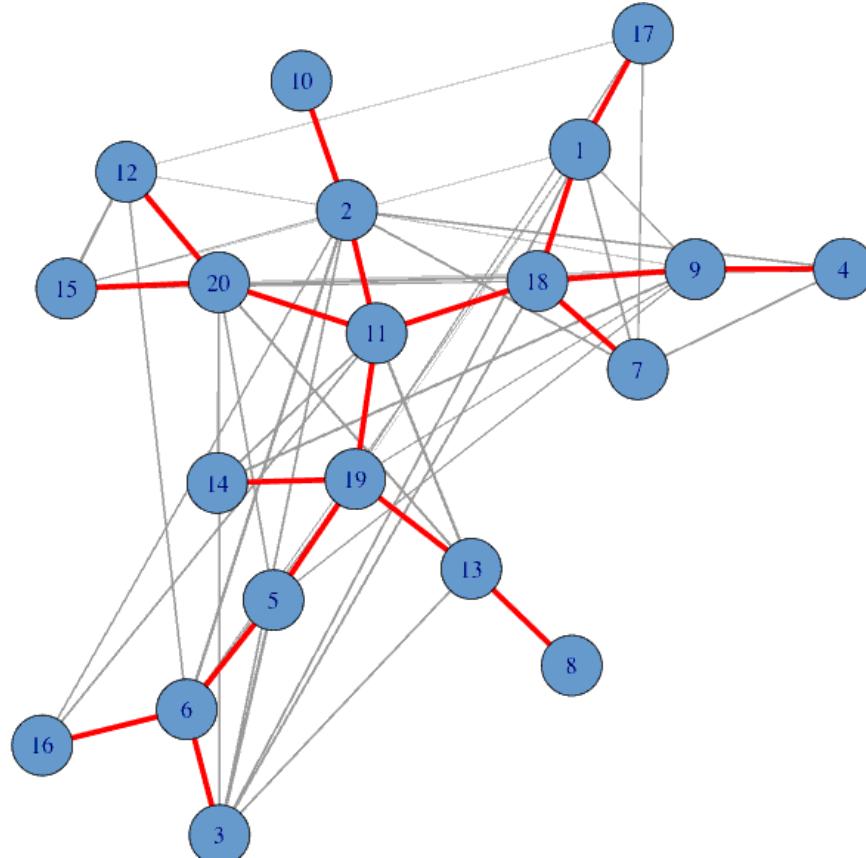
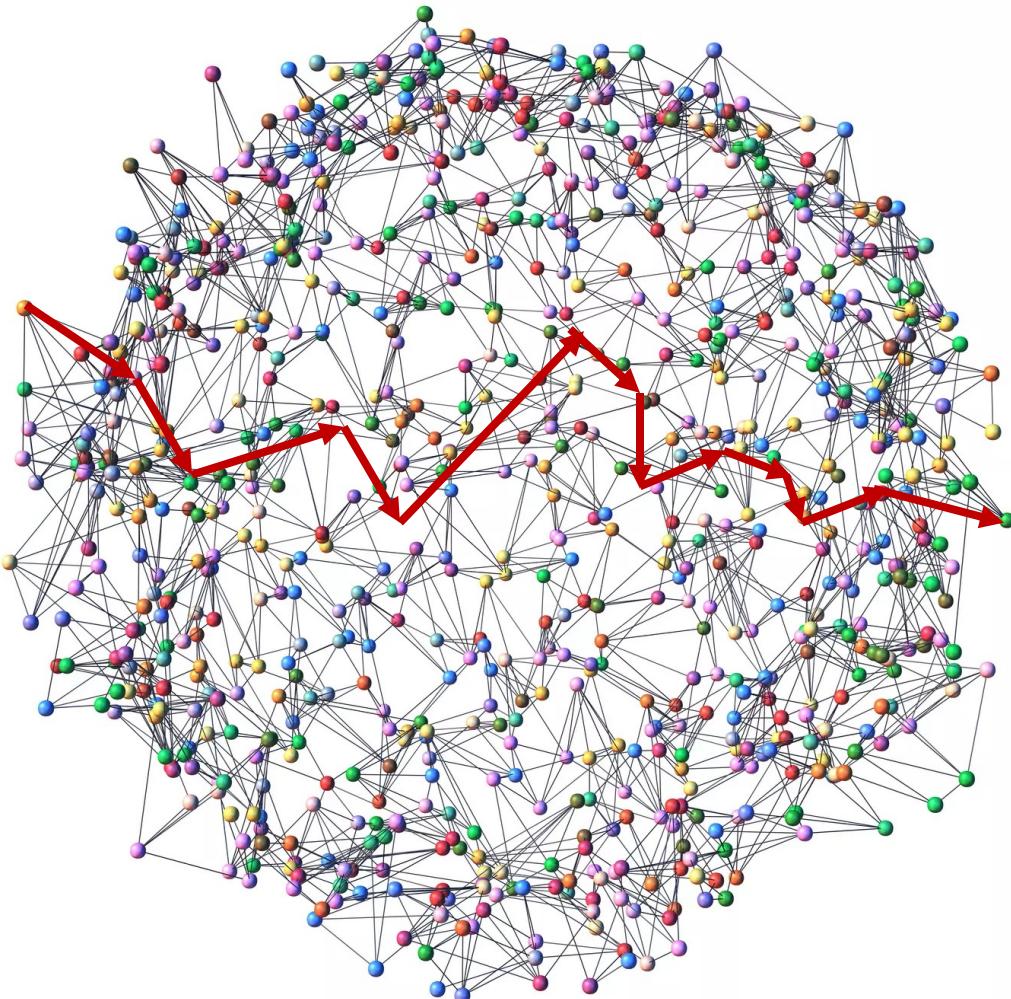
Track visited nodes
in a QUEUE



Track visited nodes
in a STACK



Graph Traversal, Finding Paths and Spanning Trees



Directory Structure

pytest	<i>Project Name</i>
dstruct	<i>package</i>
queue.py	file1.py
stack.py	file2.py
tests	<i>test-directory</i>
test_queue.py	test_file1.py
test_stack.py	test_file2.py



Running pytest

Report # of passes and failures only

\$ python -m pytest

Report outcome of each test

\$ python -m pytest -v

Report code coverage percentages (requires pytest-cov)

\$ python -m pytest -v --cov

Report coverage and identify missing lines

\$ python -m pytest -v --cov --cov-report term-missing



Running pytest

```
===== 6 passed in 0.03s =====
(base) [552 uranus:ds3500 ] python -m pytest -v --cov --cov-report term-missing
===== test session starts =====
platform darwin -- Python 3.8.8, pytest-6.2.4, py-1.10.0, pluggy-0.13.1 -- /Users/rachlin/apps/anaconda3/bin/python
cachedir: .pytest_cache
rootdir: /Users/rachlin/code/ds3500
plugins: cov-2.12.1, anyio-3.3.0
collected 6 items

ds3500/dstruct/tests/test_queue.py::test_constructor PASSED
ds3500/dstruct/tests/test_queue.py::test_enqueue PASSED
ds3500/dstruct/tests/test_queue.py::test_dequeue PASSED
ds3500/dstruct/tests/test_stack.py::test_constructor PASSED
ds3500/dstruct/tests/test_stack.py::test_push PASSED
ds3500/dstruct/tests/test_stack.py::test_pop PASSED

----- coverage: platform darwin, python 3.8.8-final-0 -----
Name          Stmts  Miss  Cover  Missing
-----
ds3500/dstruct/queue.py      17      0  100%
ds3500/dstruct/stack.py     15      0  100%
ds3500/dstruct/tests/test_queue.py  27      0  100%
ds3500/dstruct/tests/test_stack.py  21      0  100%
-----
TOTAL          80      0  100%

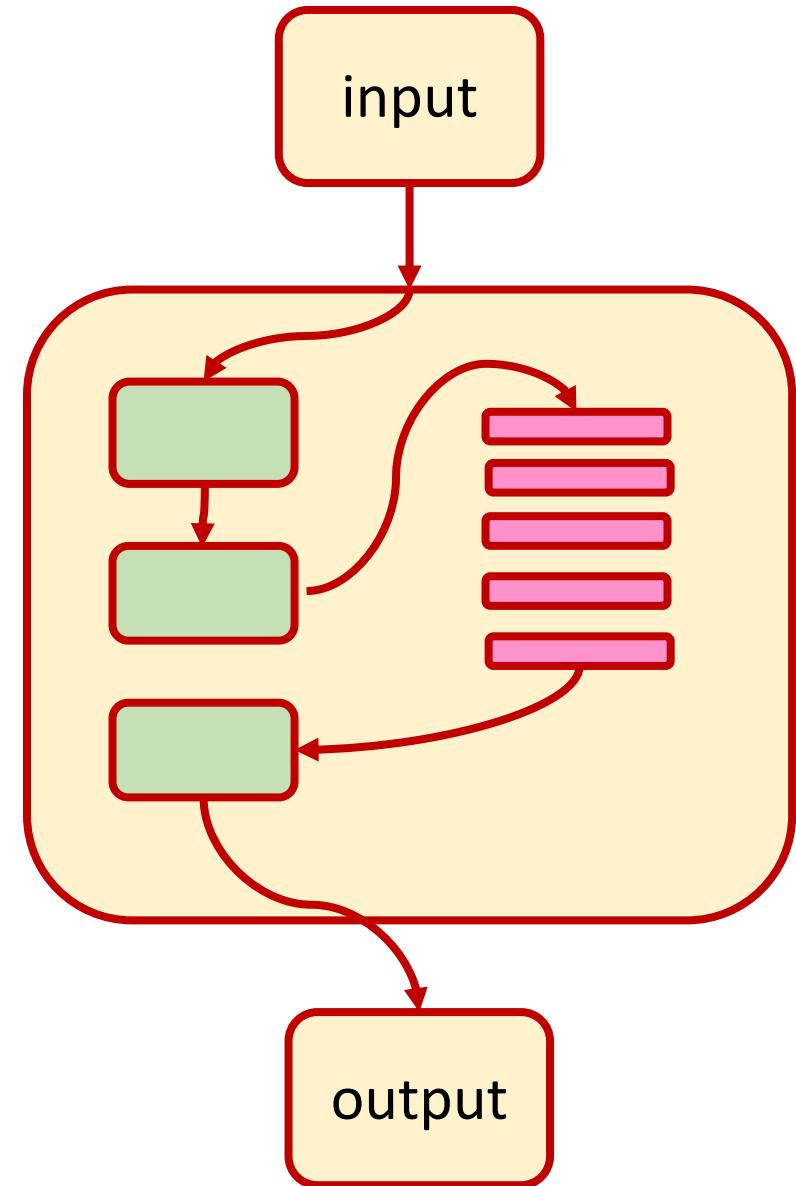
===== 6 passed in 0.09s =====
(base) [553 uranus:ds3500 ]
```



The procedural paradigm

Programs are *recipes*: a series of statements that transform our data into visualizations and insight.

In procedural programming, we manage complexity by being **modular**: adhering to top-down programming practices that break down difficult tasks into a sequence of more manageable sub-tasks.



Recipes are a powerful metaphor for data science!



start with
raw data



load
data



data refinement
(munging)



build data
structures

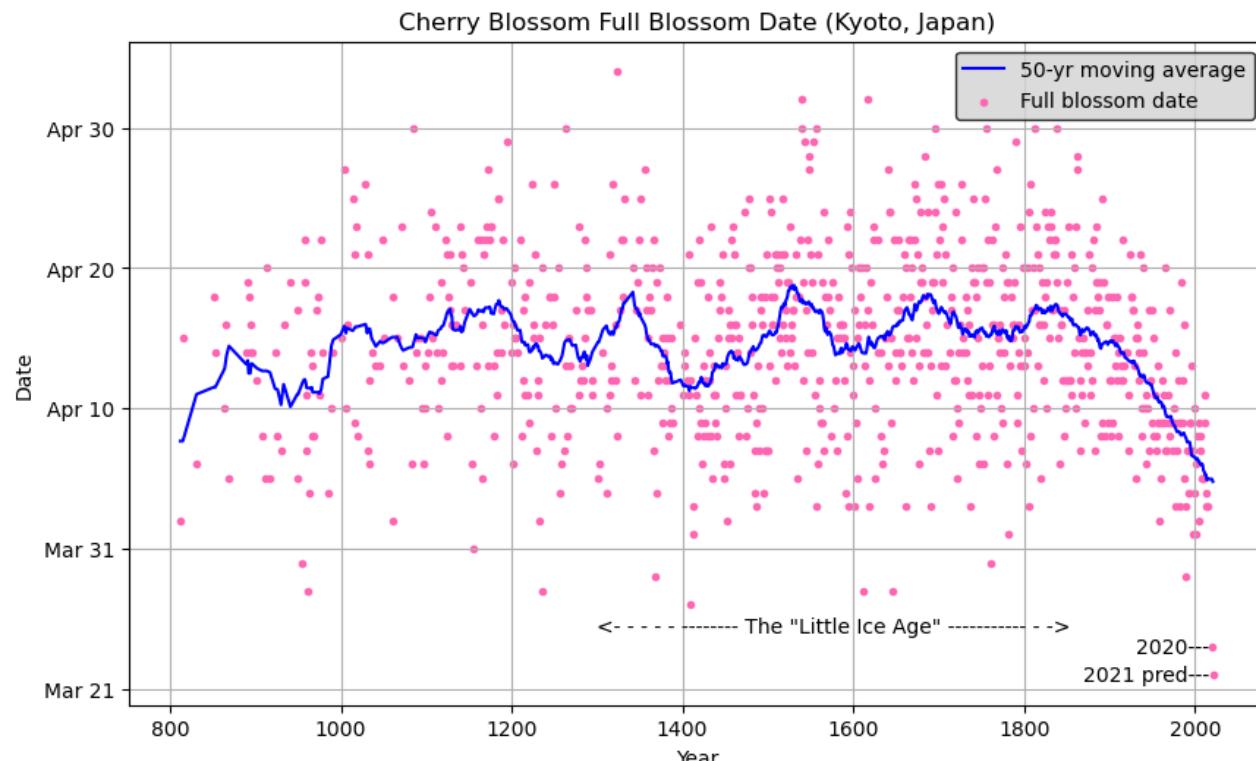
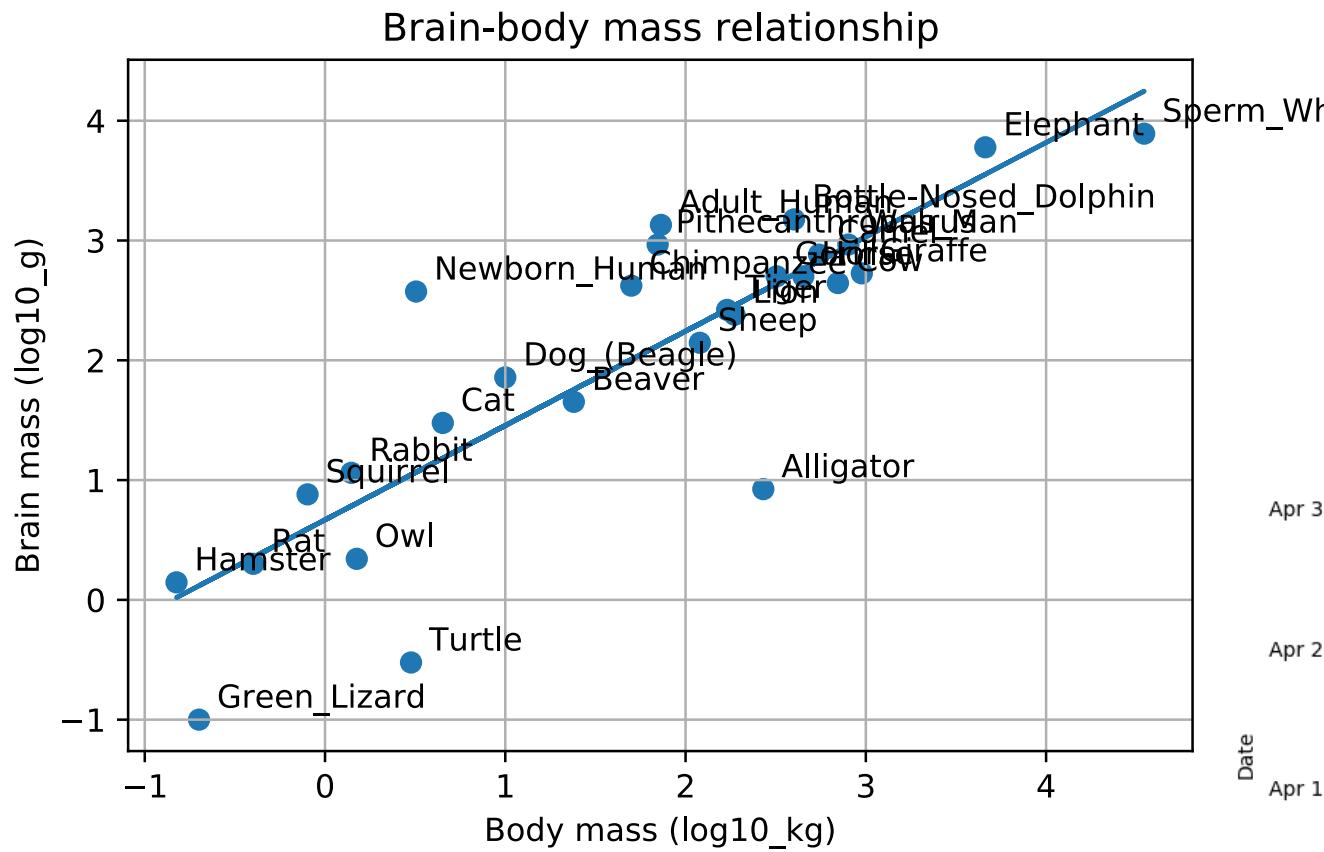


processing
and
visualization

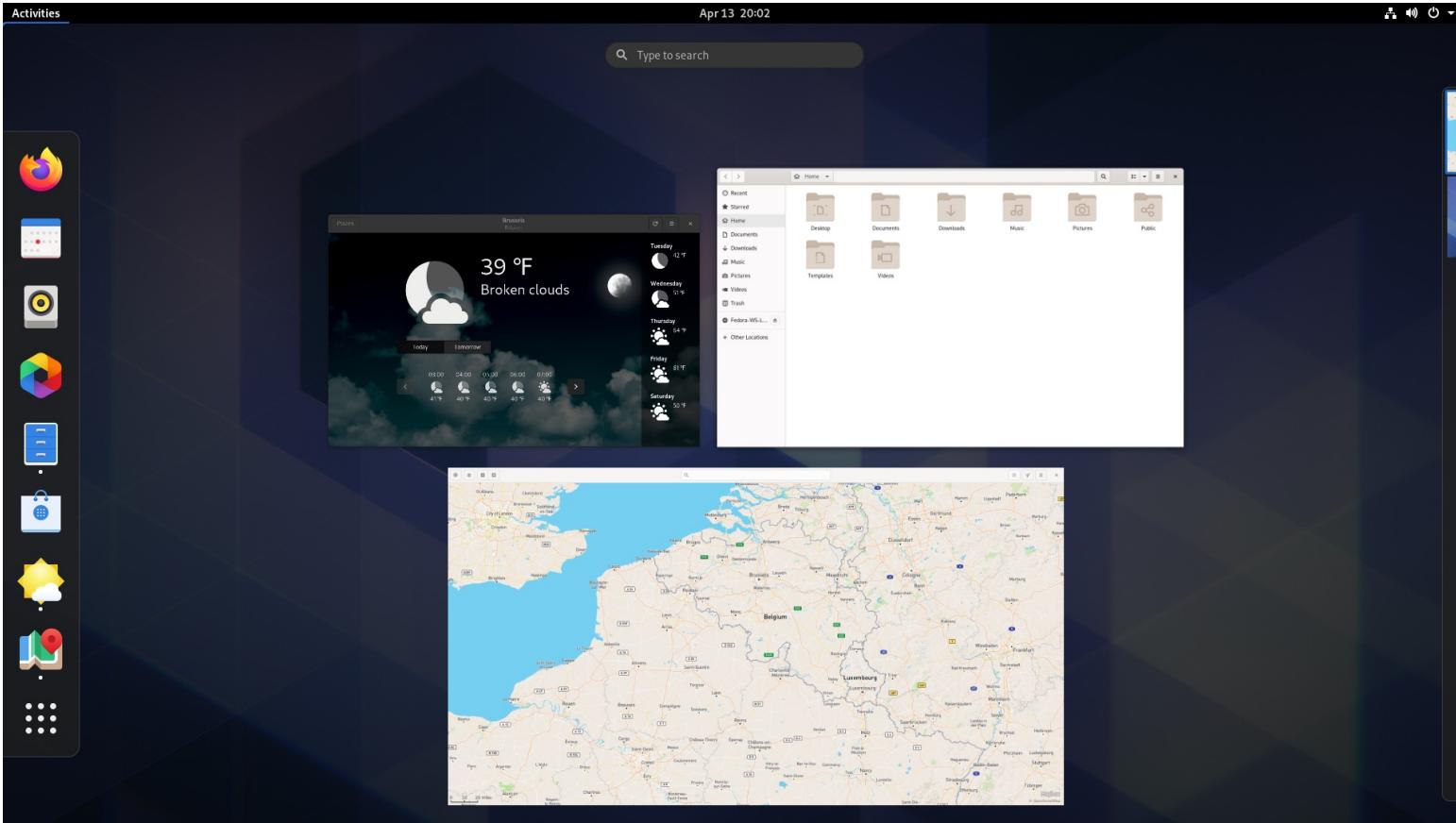


don't forget
to label your
axes!

...and it's adequate for a wide range of use-cases!



But consider the Desktop



source: <https://www.gnome.org>

We **interact** with the desktop **interface** by performing actions on **objects** that each support well-defined **behaviors**.

Objects

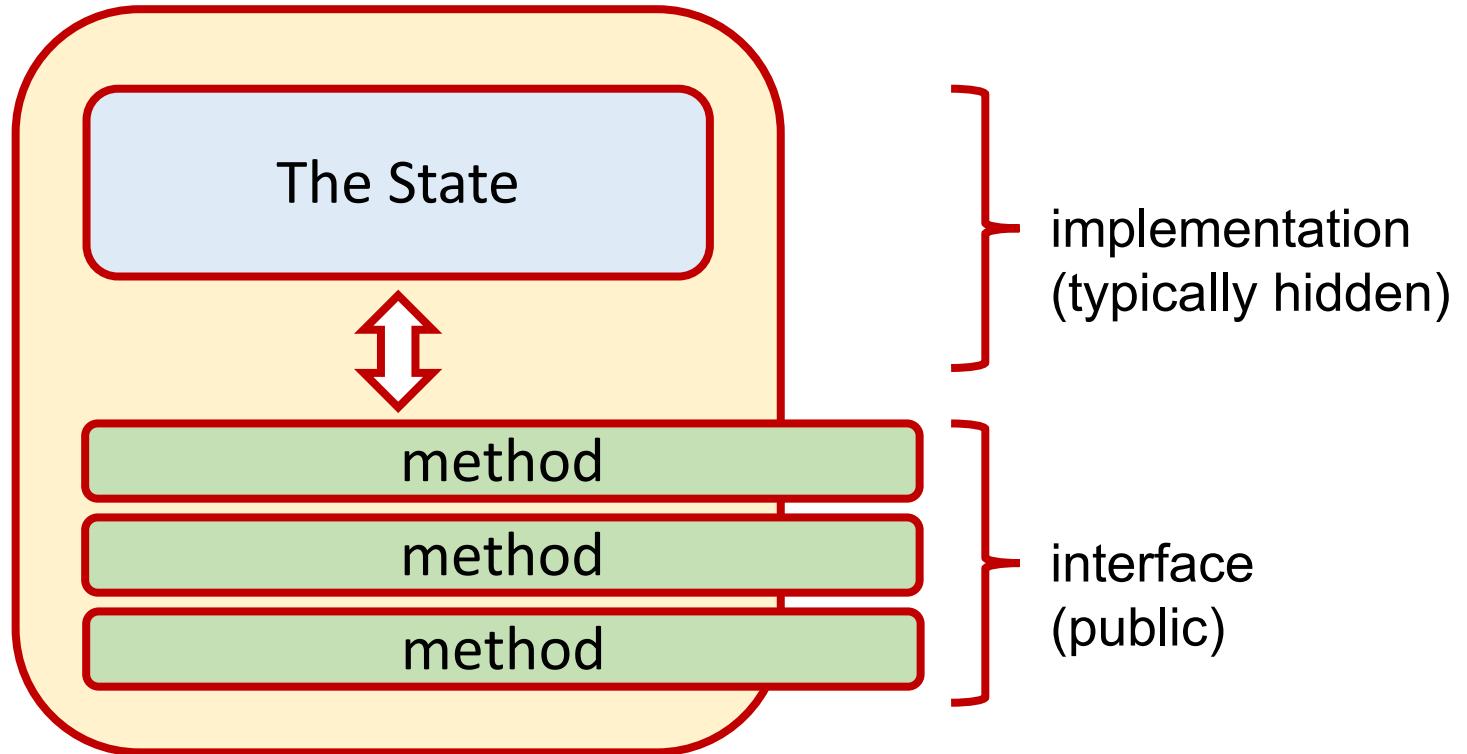
- files and folders
- windows
- apps
- menus
- status bar

Behaviors

- open/close
- move
- resize

The paradigm has clearly changed!

The object paradigm



Objects have fields and attributes that constitute the **state** of the object.

The state is accessed and modified through various **methods** that constitute the object's **interface**.

The layout and organization of the state (i.e., the **implementation**) is usually shielded from the user.

The Interface/Implementation Dichotomy



di·chot·o·my

/dī'kădəmē/

noun

a division or contrast between two things that are or are represented as being opposed or entirely different.

"a rigid **dichotomy** between science and mysticism"

Similar: [division](#) [separation](#) [divorce](#) [split](#) [gulf](#) [chasm](#) [difference](#) ▾

We can operate complex machines without knowing what's *under the hood*.

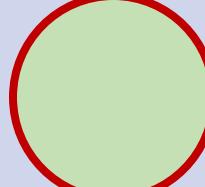
Similarly, to build more complex software, we need to express ideas at a higher level of abstraction with a focus of interface over implementation. Object-oriented thinking enables us to do this.



Classes and Objects

Classes define a **type**. It acts as a *template* or blueprint.

We then construct many objects that are **instances** of a particular class.

Class (Type)	List	Circle	Cat	Account
Object Instances	[1, 2, 3] ['Jack', 'Abby'] [] [(0,0), (5,7), (-2,2)]	 	 	my_cat not_my_cat



Lists, Tuples, and Dictionaries are objects!

```
L = ['A', 'B', 'C', 'D']
```

```
# What is the class/type?
```

```
type(L)  
list
```

```
# Call a method on the object
```

```
L.append('E')
```

```
L  
['A', 'B', 'C', 'D', 'E']
```

```
# Pass L to a function – we can still do this!
```

```
len(L)  
5
```

```
L[2]  
'C'
```

```
T = ('A', 'B', 'C', 'D')
```

```
type(T)  
tuple
```

```
len(T)  
4
```

```
T[2]  
'C'
```

```
D = {'ann':44, 'reuban':29, 'dachuan':37}
```

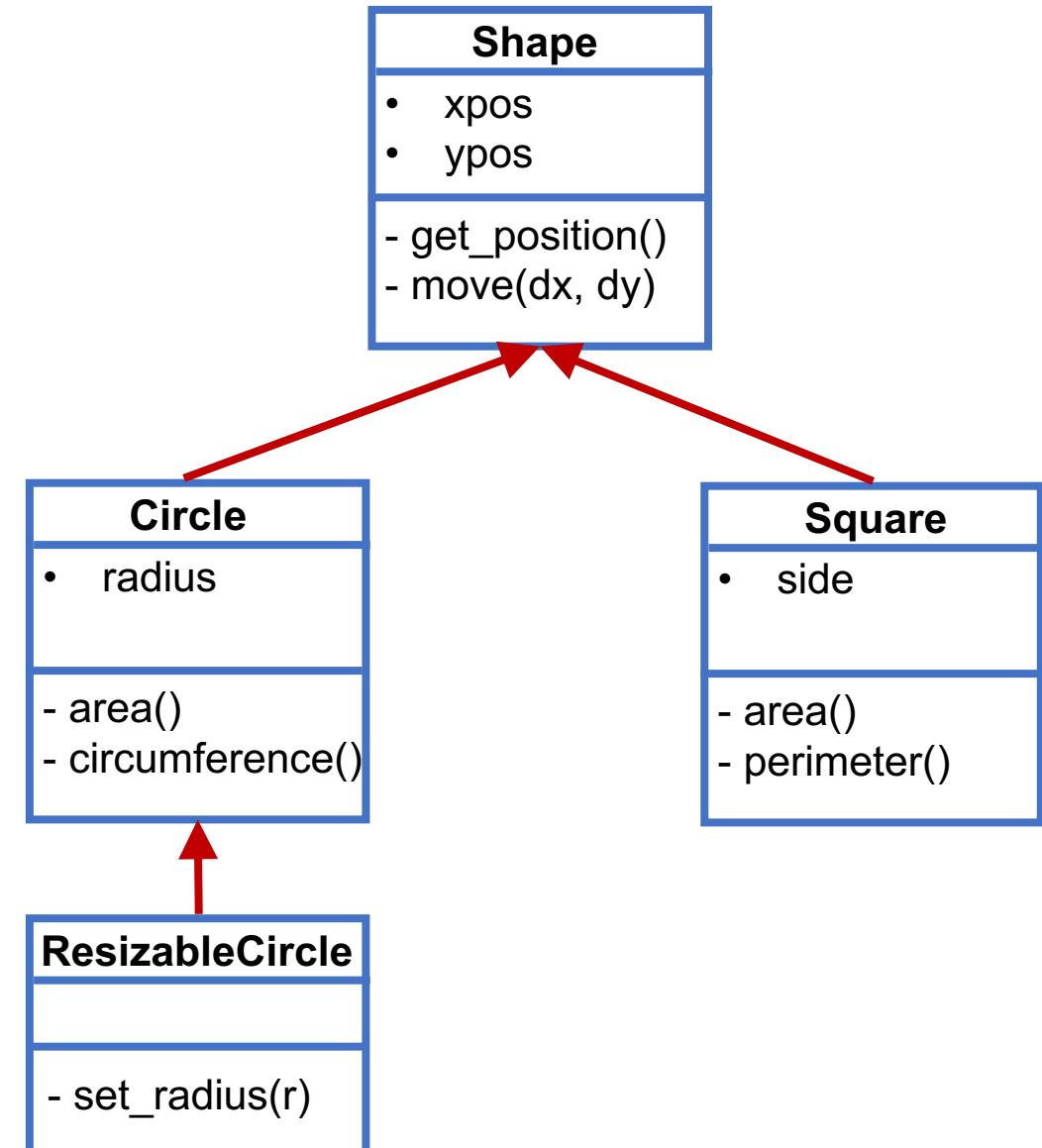
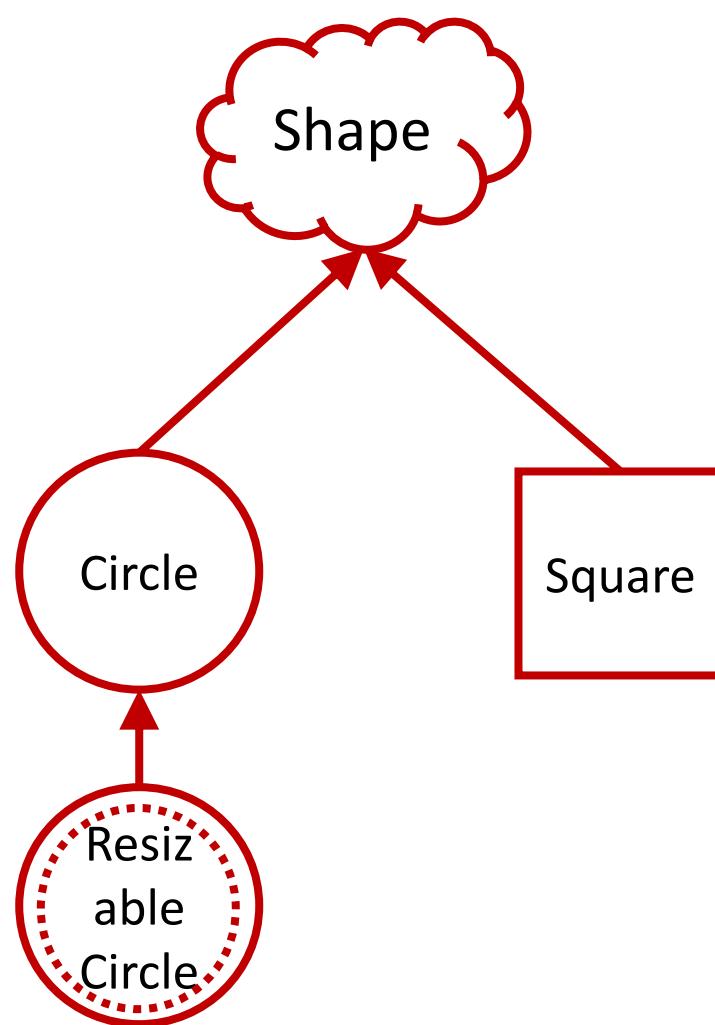
```
D['reuban']  
29
```

```
len(D)  
3
```

```
list(D.keys())  
['ann', 'reuban', 'dachuan']
```



Class Inheritance



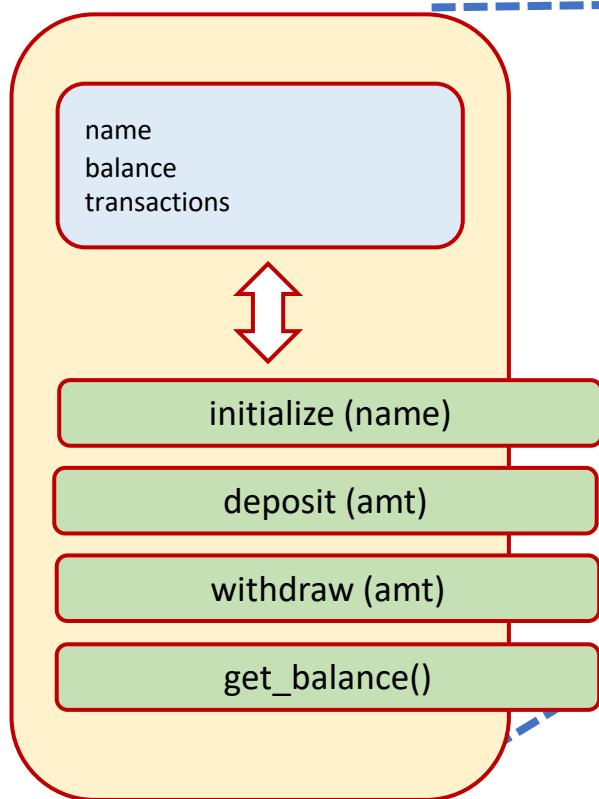
Abstract Base Classes (ABCs)

Abstract base classes define a set of methods and properties that a class must implement in order to be considered a duck-type instance of that class.

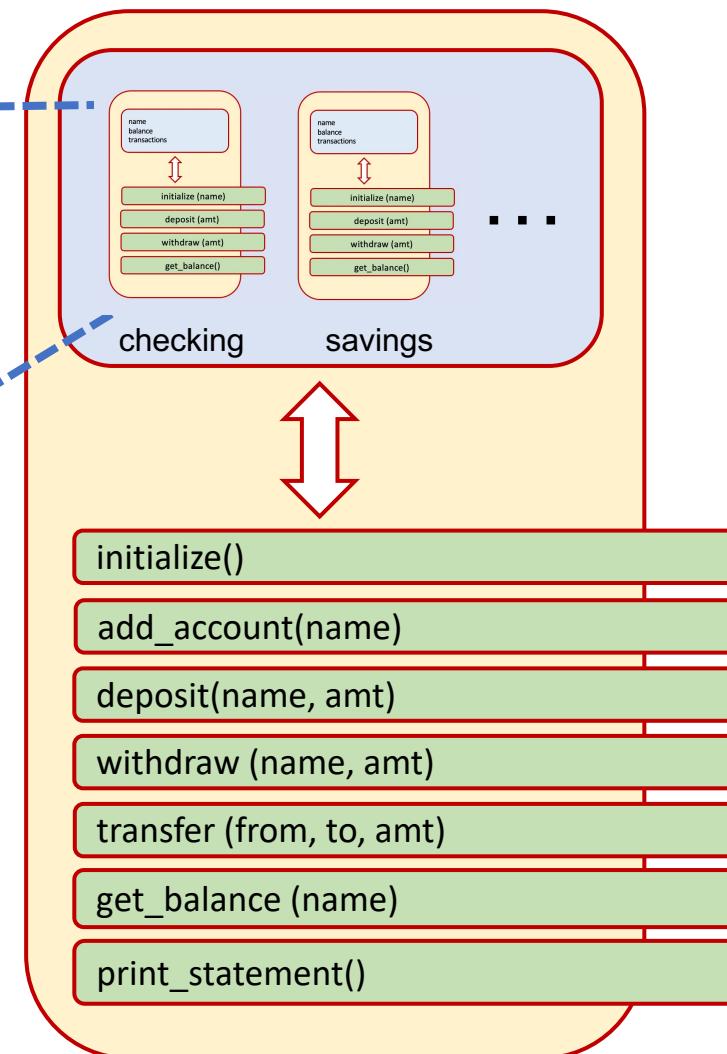


Composition

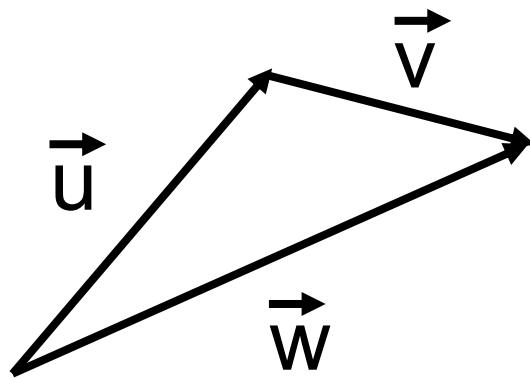
Account



Banking



Operator overloading



This:

```
u = Vector(3, 4)
```

```
v = Vector(2, -1)
```

```
w = u + v
```

```
w  
5i + 3j
```

```
u * v # dot product  
2
```

Not this:

```
u = [3, 4]
```

```
v = [2, -1]
```

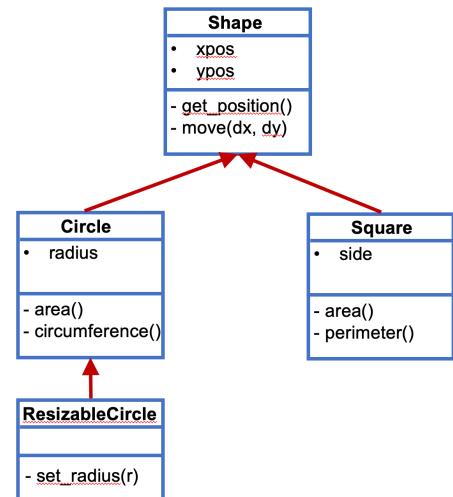
```
w = add(u, v)
```

```
w  
[5,3]
```

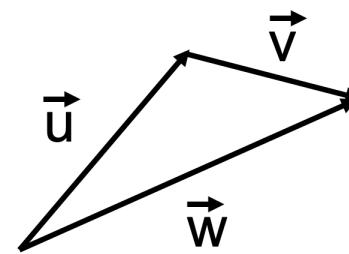
```
dot_product(u, v)  
2
```

The advantages of object-oriented design

Code re-use
(inheritance)

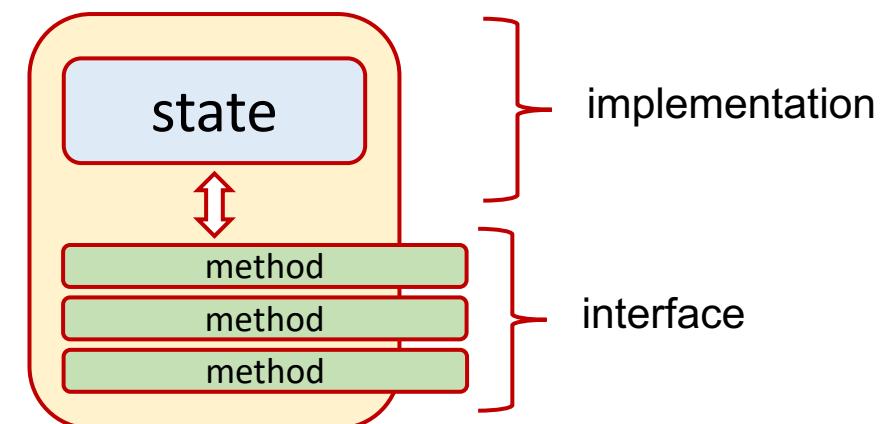


Expressive syntax
(operator overloading)



$$w = u + v$$

Managing complexity
(separation of implementation and interface)



Object-Oriented Programming

Design Patterns



Northeastern University

Design Patterns

Design Patterns: Elements of Reusable Object-Oriented Software 1st Edition

by [Erich Gamma](#) ▾ (Author), [Richard Helm](#) ▾ (Author), [Ralph Johnson](#) ▾ (Author), [John Vlissides](#) ▾ (Author), [Grady Booch](#) (Foreword)

★★★★★ ▾ 1,548 ratings

#1 Best Seller in Software Reuse

Look inside ↓



ISBN-13: 978-0201633610

ISBN-10: 0201633612

[Why is ISBN important?](#) ▾

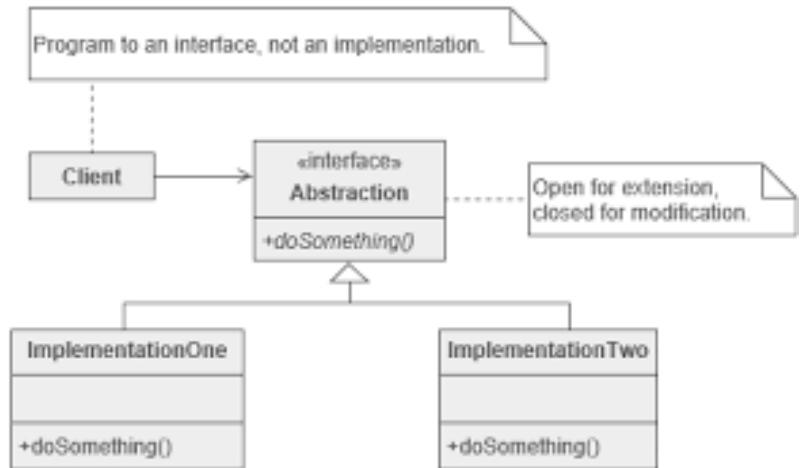
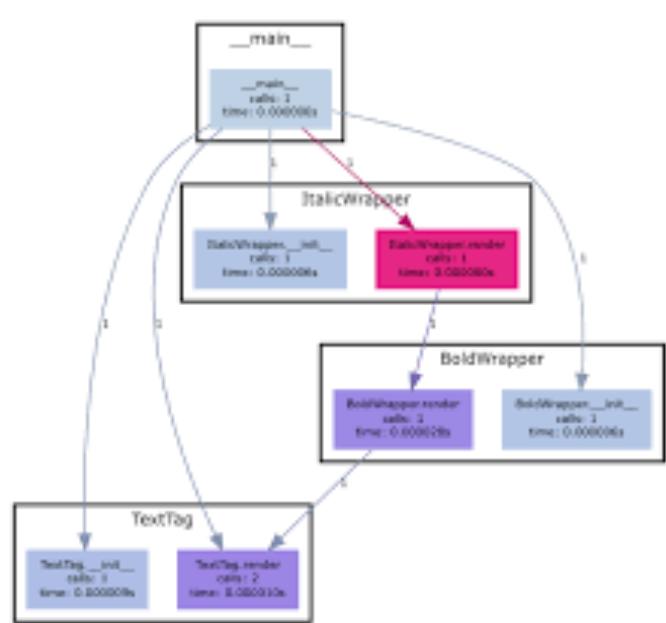
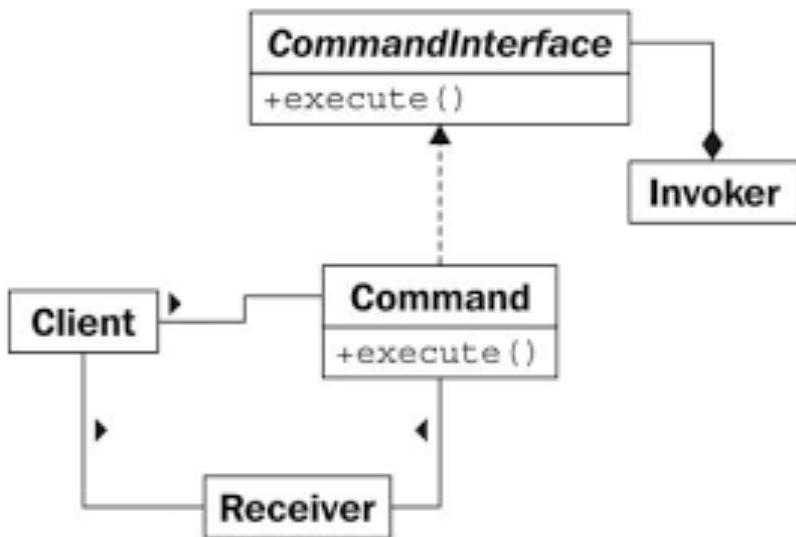
Kindle	Hardcover	Paperback	Other Sellers
\$32.39	\$27.99 - \$37.85	from \$32.97	See all 12 versions
<input type="radio"/> Buy used:		\$27.99	
<input checked="" type="radio"/> Buy new:		\$37.85	
In Stock.		List Price: \$59.99 Details Save: \$22.14 (37%) 	
Ships from and sold by Amazon.com.		FREE delivery Tomorrow, October 15. Order within 4 hrs 32 mins	
May be available at a lower price from other sellers , potentially without free Prime shipping.		Deliver to John - Fountain Hills 85268	
		Qty: 1 ▾	Add to Cart
			Buy Now



Northeastern University

Design Pattern Objectives

Common software problems should be solved in standard ways.
(Don't re-invent the wheel.)



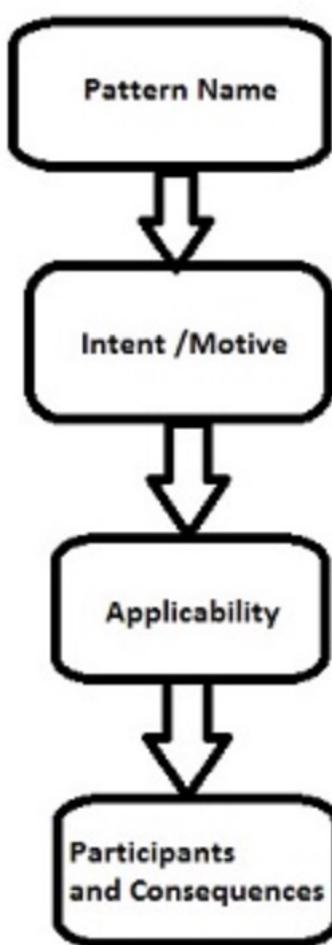
What is a Design Pattern?

- In software engineering, a **software design pattern** is a general, reusable solution to a commonly occurring problem within a given context in software design.
- It is not a finished design that can be transformed directly into source or machine code. Rather, it is a description or template for **how to solve a problem** that can be used in many different situations.

Design patterns are **formalized best practices** that the programmer can use to solve common problems when designing an application or system.



Design Patterns



Pattern Name

It describes the pattern in short and effective manner.



Intent/Motive

It describes what the pattern does.



Applicability

It describes the list of situations where pattern is applicable.



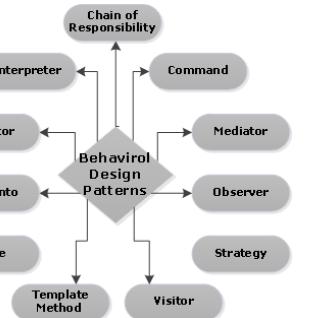
Participants and consequences

Participants include classes and objects that participate in the design pattern with a list of consequences that exist with the pattern.

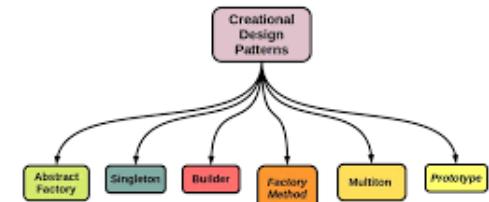


Types of Design Patterns

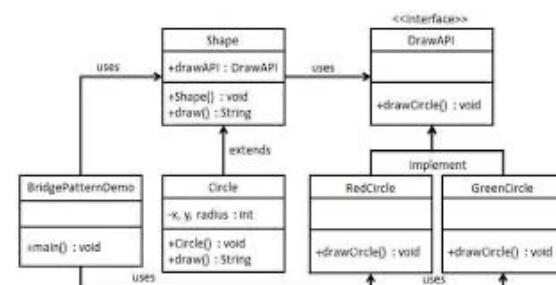
1. **Behavioral patterns** describe interactions between objects and focus on how objects communicate with each other. Behavioral patterns are also used to make the algorithm that a class uses simply another parameter that is adjustable at runtime.



2. **Creational patterns** are used to create objects for a suitable class that serves as a solution for a problem.

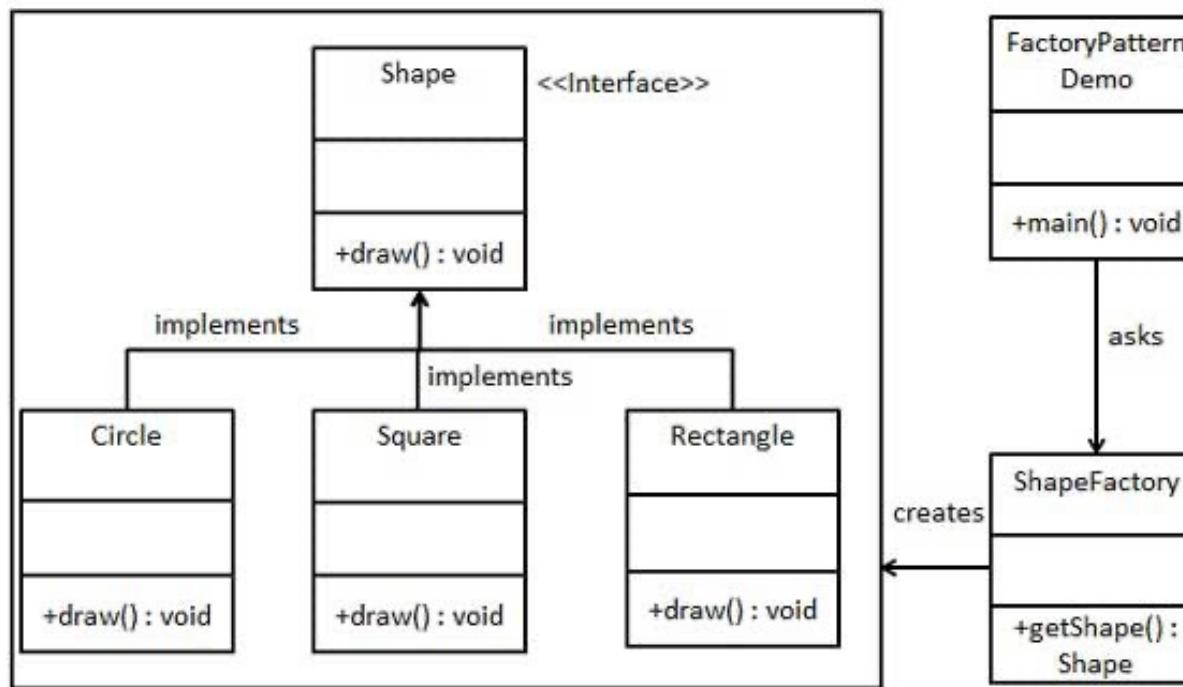


3. **Structural patterns** form larger structures from individual parts, generally of different classes. Structural patterns vary a great deal depending on what sort of structure is being created for what purpose.



Factory Method (A Creational Design Pattern)

- Create a concrete implementation of a common interface
- Separate creation process from the code that depends on the interface
- Delegate the creation of objects to the factory



When to use a Factory Pattern

18 Answers

Sorted by: Highest score (default) 

 I like thinking about design patterns in terms of my classes being 'people,' and the patterns are the ways that the people talk to each other.

419

 So, to me the factory pattern is like a hiring agency. You've got someone that will need a variable number of workers. This person may know some info they need in the people they hire, but that's it.

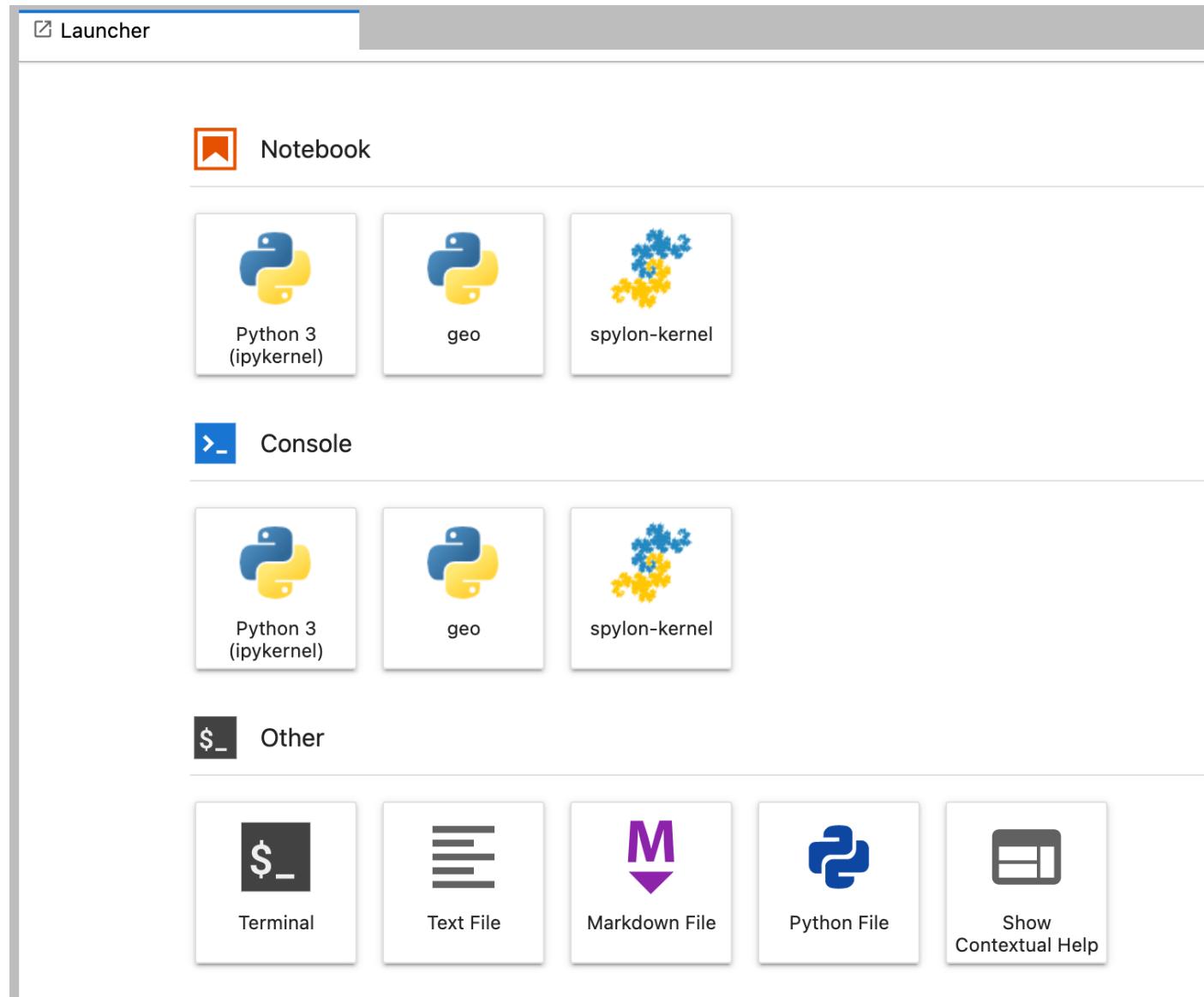


So, when they need a new employee, they call the hiring agency and tell them what they need. Now, to actually *hire* someone, you need to know a lot of stuff - benefits, eligibility verification, etc. But the person hiring doesn't need to know any of this - the hiring agency handles all of that.

In the same way, using a Factory allows the consumer to create new objects without having to know the details of how they're created, or what their dependencies are - they only have to give the information they actually want.



The Jupyter Lab Launcher is a Factory



SOLID PRINCIPLES

S

SINGLE
RESPONSIBILITY

A class should have only single responsibility and should have one and only one reason for change

O

OPEN CLOSED
PRINCIPLE

A class should be open for extension, but closed for modifications

L

LISKOV
SUBSTITUTION

Objects in a program should be replaceable with instances of their subtypes without altering the correctness of program

I

INTERFACE
SEGREGATION

Segregate Interfaces as per the requirements of program, rather than one general purpose implementation

D

DEPENDENCY
INVERSION

Should depend on abstractions rather than concrete implementations

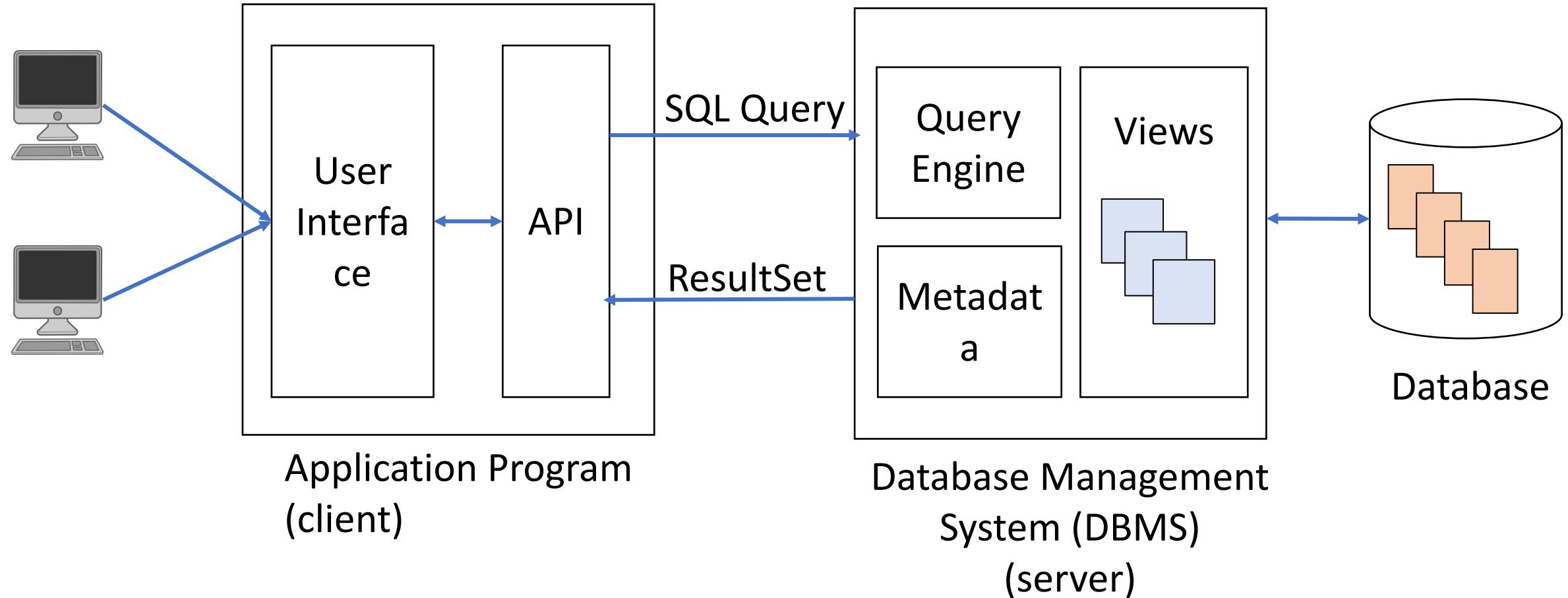


SRP

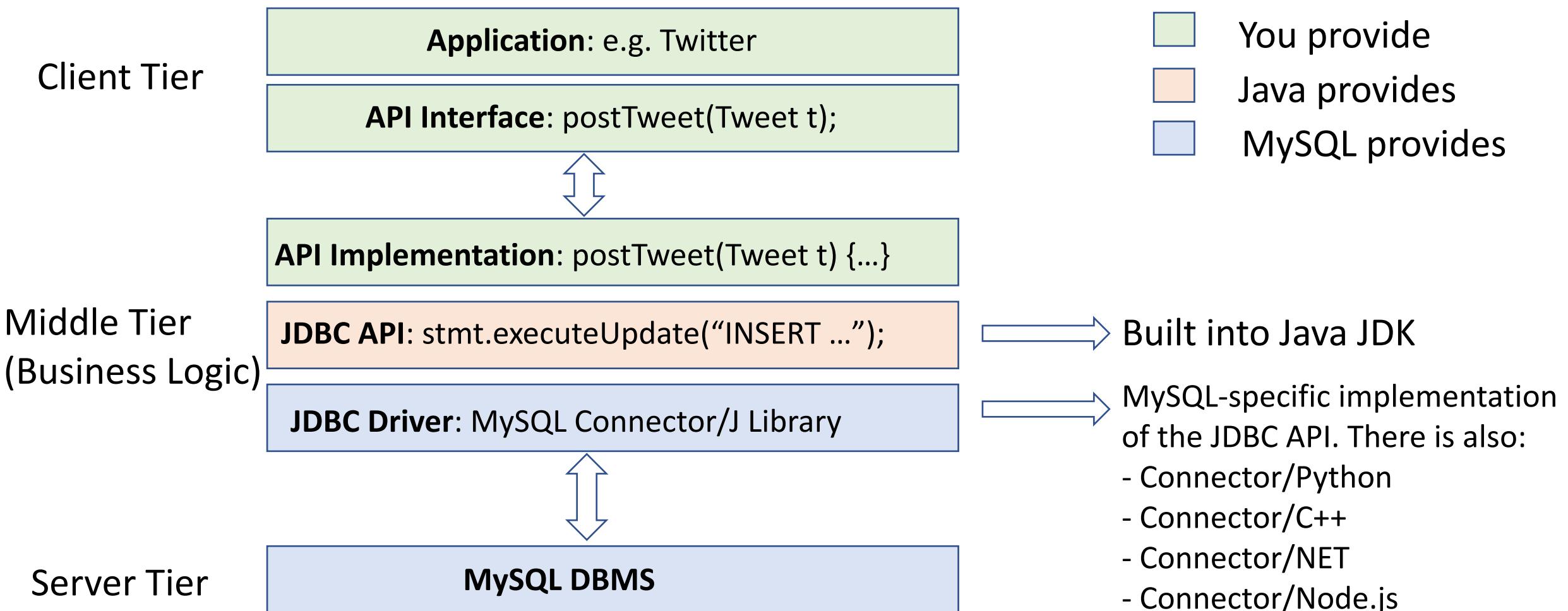
The **single-responsibility principle (SRP)** is a computer-programming principle that states that every module, class or function in a computer program should have responsibility over a single part of that program's functionality, and it should encapsulate that part. All of that module, class or function's services should be narrowly aligned with that responsibility. - *Wikipedia*



Two-Tier Client-Server Design Pattern



Client-Server Design Pattern (Twitter - Java)



Client-Server Design Pattern (BioApplication - Python)

Client Tier
(Application)

class: BioApplication (bioapp.py)
method: local_phenotype_network("asthma");



You provide



MySQL provides

Middle Tier
(DB Interface)

dbutils.py:
class: DBUtils → execute(query)
class: ConnectionFactory → get_connection()

Driver: MySQL Connector/Python Library
method: connect / close
method: cursor / execute / fetchall



Python-specific implementation
of the MySQL Driver. There is also:
- Connector/J
- Connector/C++
- Connector/.NET
- Connector/Node.js

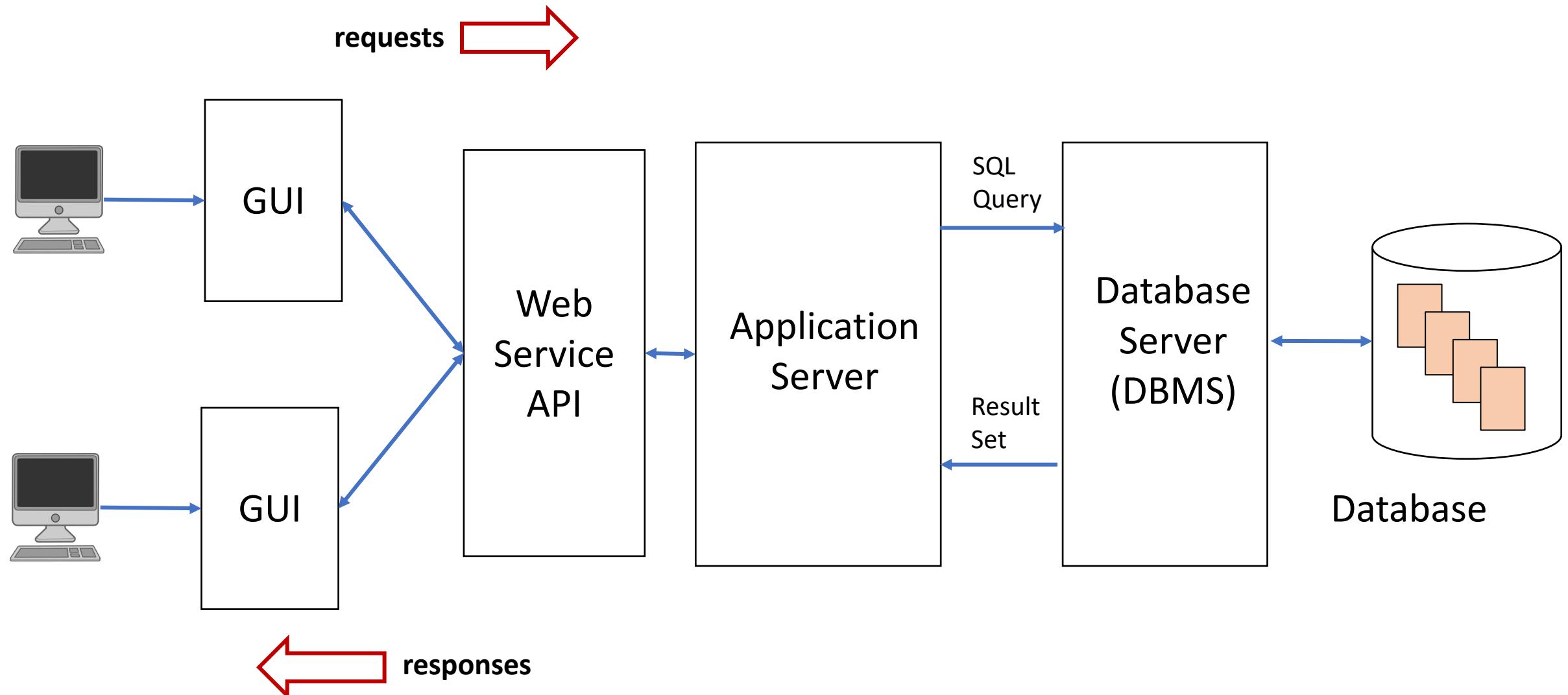
Server Tier

MySQL DBMS



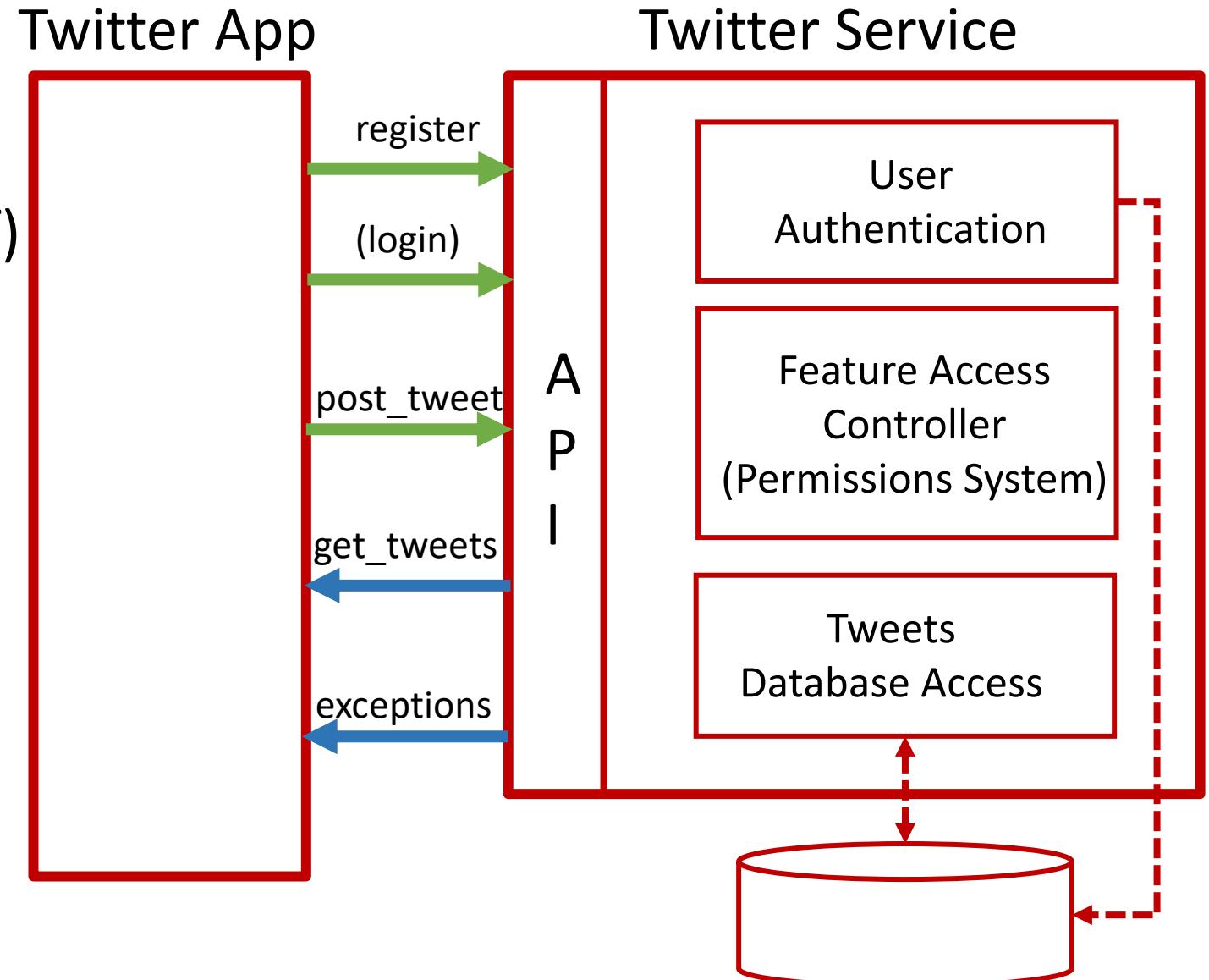
Northeastern University

Three-Tier Client-Server Design Pattern



Monolith Twitter

- **User authentication:** responsible for validating user credentials (login / logoff)
- **Permission System:** Keeps track of what features users have access to.
- **Tweets database:** Stores and retrieves user tweets

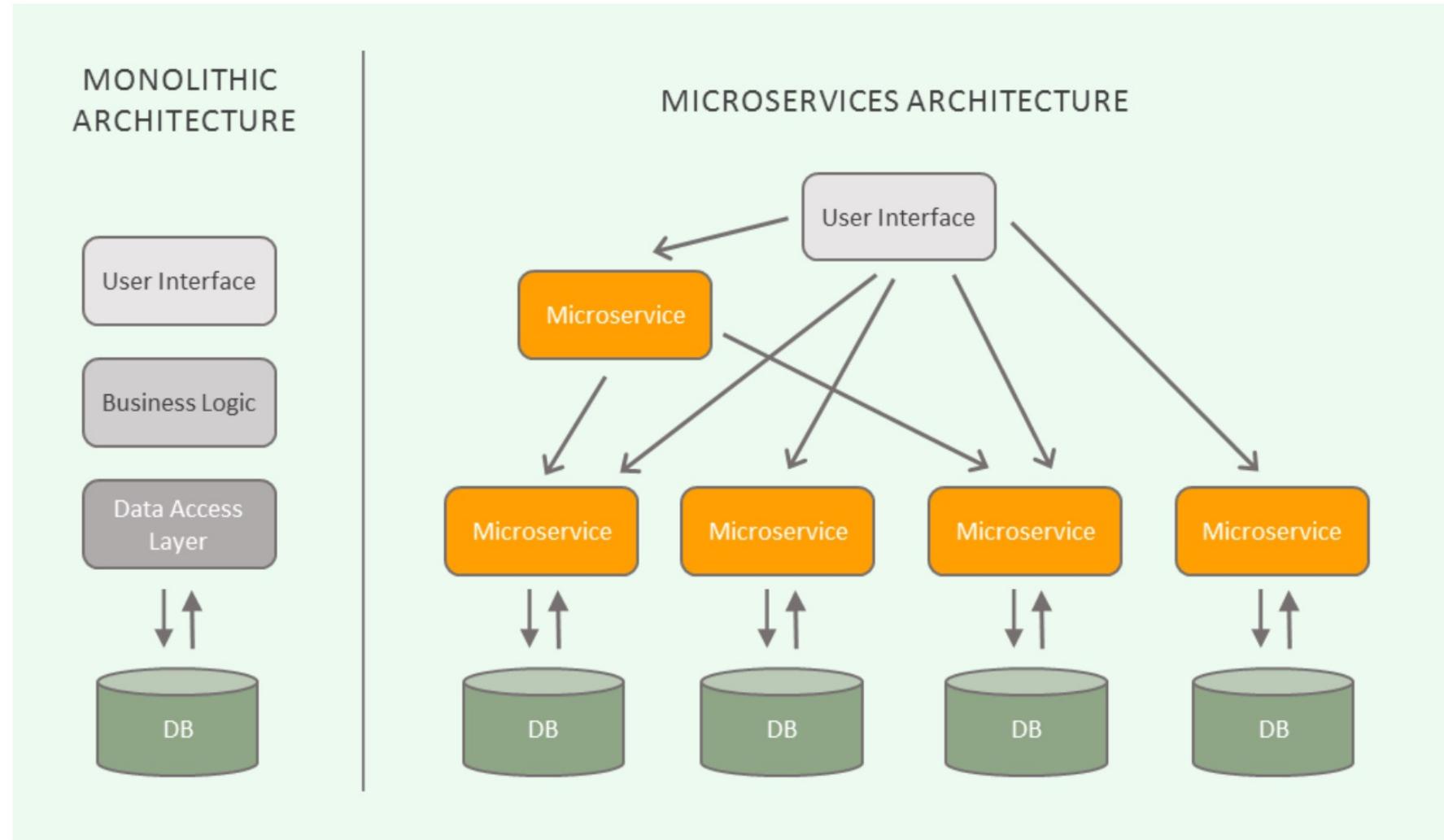


Disadvantages of Monolithic Architectures

- The application codebase can become very large and difficult to understand by any one person
- Difficult to maintain – requires careful testing at both the component and application level
- Longer deployment times



Monolith vs. Microservices



Source: DZone

Northeastern University



Microservices have their challenges too!

- As a loosely-coupled distributed system, microservices tend to be more complex – there are simply more moving pieces that need to be setup, tested, and monitored.
- Failures due to cross-service communication delays: Developers must address problems such as network latency and load balancing
- Integration testing is more challenging. Microservices are siloed. So are the developers that manage the individual services.



Hacker News (news.ycombinator.com) takes up the debate

Response to the blob post: ***Give me back my monolith***

<https://www.craigkerstiens.com/2019/03/13/give-me-back-my-monolith/>

- **Ease of setup:**

I'm not sure why you would start with microservices, unless you wanted to show that you could build a microservices application. Monoliths are quicker and easier to setup when you're talking about a small service in the first place.

- **Scalability considerations:**

It's when an organization grows and the software grows and the monolith starts to get unwieldy that it makes sense to go to microservices. It's then that the advantage of microservices both at the engineering and organizational level really helps.

- **When microservices really shine**

A team of three engineers orchestrating 25 microservices sounds insane to me.

A team of thirty turning one monolith into 10 microservices and splitting into 10 teams of three, each responsible for maintaining one service, is the scenario you want for microservices.



Monolith Twitter Learning Goals

- Composition of Objects to build higher-level components
- **Flexible** but **uniform** approach to context-sensitive exception handling in a non-trivial application architecture
- DO NOT STORE PASSWORDS!

What could possibly go wrong??

registration

- UsernameAlreadyExists
- InvalidPassword

authentication

- UnknownUsername
- IncorrectPassword

access control

- PermissionDenied



Typing in Python

Duck Typing, Dynamic Typing, and Type Hints



Static typing: Data types are pre-declared (Java)

- In most programming languages, the type of a variable, an object instance, a function parameter, and even the type of the function's return value (if any) must be declared and is **FIXED (static)**.
- These constraints allow the compiler to check our program *before running the code*.
- (Usually this is done at *compile time* – such languages have a compiler.)



Static Typing: continued

In Java, we might say:

```
void drawShape(Shape sh) {  
    .  
    .  
}
```

The function can accept any object instance of a Shape (*or any Shape sub-class*). We may not know the shape until runtime. This is called **runtime polymorphism**. (But it must be some kind of shape in accordance with our class hierarchy.)



Dynamic typing: Data types are determined at run-time

- In python, the interpreter figures out and checks the type at runtime. This is known as **dynamic typing**.
- Functions and methods don't specify return types or parameter types.

```
def draw_shape(sh):  
    """ Draw a shape """  
    sh.draw()  
  
    .  
    .
```



Dynamic Typing

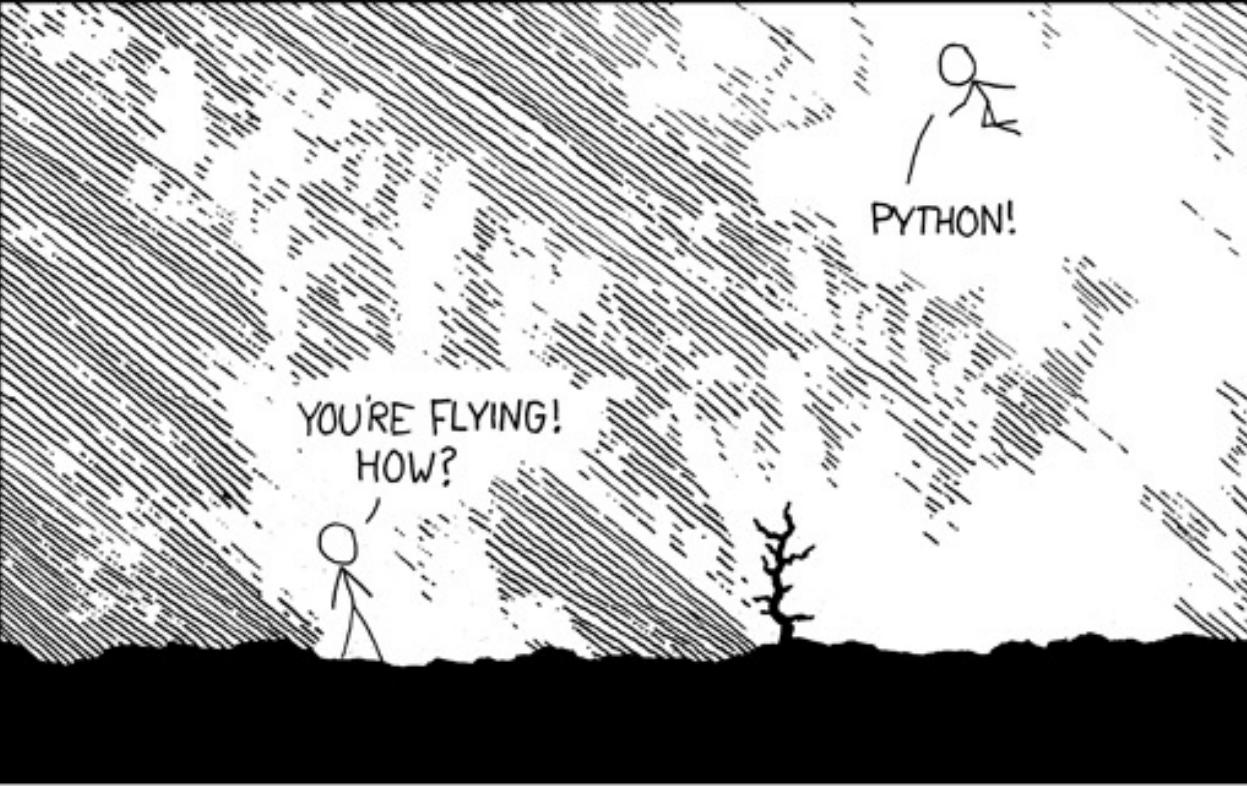
Advantage: Greater flexibility

Disadvantage:

- We don't know there is a problem until the program crashes.
- The absence of a compiler means that certain code optimizations can not be performed.
- Interpreted languages like python will tend to run slower!

*How do we know that **sh** is part of the Shape class hierarchy that support a **draw** method? (We don't)*





I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!

HELLO WORLD IS JUST
`print("Hello, world!")`

I DUNNO...
DYNAMIC TYPING?
WHITESPACE?

COME JOIN US!
PROGRAMMING IS FUN AGAIN!
IT'S A WHOLE NEW WORLD UP HERE!

BUT HOW ARE YOU FLYING?

I JUST TYPED
`import antigravity`

THAT'S IT?

... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.

BUT I THINK THIS IS THE PYTHON.



Type Hints

Type hints make it possible to do static typing (with the help of other tools). They are not enforced!

Introduced in PEP 484 / Python version 3.5+:

```
def greeting(name: str) -> str:  
    return 'Hello ' + name
```

From the PEP: *Python will remain a dynamically typed language, and the authors have no desire to ever make type hints mandatory, even by convention.*



Type Hints in PyCharm

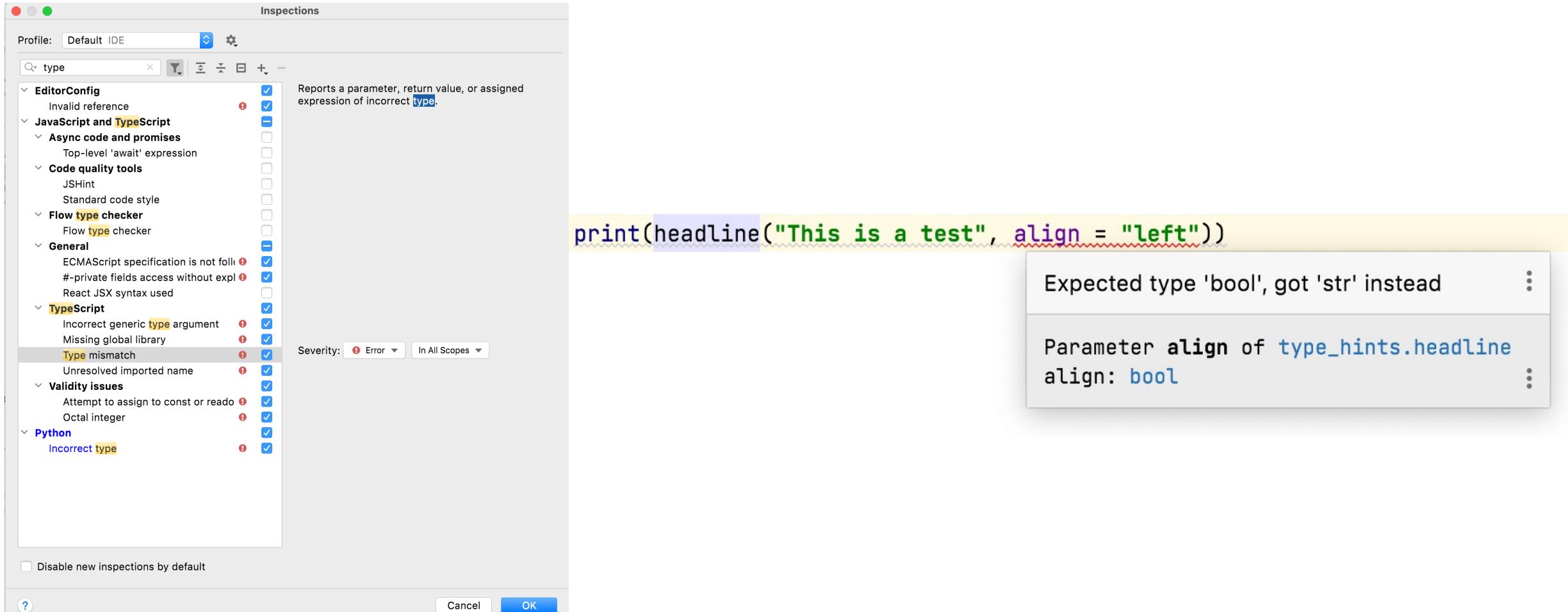
A 1 A 5 ^ v

```
def headline(text: str, align: bool = True) -> str:  
    if align:  
        return f"{text.title()}\n{'-'* len(text)}"  
    else:  
        return f" {text.title()} ".center(40, "*")  
  
print(headline("This is a test", align = "center"))
```

⚠ Expected type 'bool', got 'str' instead :9



Type Hints in PyCharm can be converted to Errors



Type Hint Conventions

PEP 8 Recommends

- **Use normal rules for colons, no space before and one space after a colon**

text: str

- **Use spaces around the = sign when combining an argument annotation with a default value**

align: bool = True

- **Use spaces around the -> arrow**

def headline(...) -> str



Mypy (mypy-lang.org): A static type checker for python

[Home](#) [Blog](#) [Examples](#) [Docs](#) [GitHub](#) [Wiki](#) [Roadmap](#) [About](#) [Contact](#) [Newsletter](#) [Twitter](#)

Follow: [Atom](#)

mypy

Mypy is an optional static type checker for Python that aims to combine the benefits of dynamic (or "duck") typing and static typing. Mypy combines the expressive power and convenience of Python with a powerful type system and compile-time type checking. Mypy type checks standard Python programs; run them using any Python VM with basically no runtime overhead.

[Get the Code on GitHub!](#)



Northeastern University

Mypy (mypy-lang.org): A static type checker for python

```
(base) [522 uranus:10_patterns1 ] mypy type_hints.py
type_hints.py:8: error: Argument "align" to "headline" has incompatible
type "str"; expected "bool"
Found 1 error in 1 file (checked 1 source file)
(base) [523 uranus:10_patterns1 ]
```

(running mypy from the command line.)

```
(base) [524 uranus:10_patterns1 ] mypy type_hints.py
Success: no issues found in 1 source file
(base) [525 uranus:10_patterns1 ]
```



Type Hints: Advantages and Disadvantages

ADVANTAGES:

- Extra pre-runtime checking
- Improved readability / documentation – better than docstrings
- Improved functionality in certain IDEs like PyCharm
- Improved code / architecture
- Improved reliability for published libraries

DISADVANTAGES:

- More verbose code
- Limited to Python 3+
- Advanced features require 3.6+
- Slightly slower program start-up
- Not useful for small throw-away code



Type Hints are a specific kind of Python Annotation

Annotations such as Type Hints can be *inspected* programmatically at runtime using the

- `__annotations__`

attribute or method

```
{'radius': <class 'float'>, 'return': <class 'float'>}  
{'pi': <class 'float'>, 'nothing': <class 'str'>}
```



Type comments (older - used by some checkers)

```
>>> import math
>>>
>>> def circumference(radius):
...     # type: (float) -> float
...     return 2 * math.pi * radius
...
>>>
>>> circumference(4.5)
28.274333882308138
>>> █
```

Not stored in `__annotations__` but Mypy will pick them up.



Type comments (older - used by some checkers)

```
2
3 def headline1(text, width=80, fill_char="-"):
4     # type: (str, int, str) -> str
5     return f" {text.title()} ".center(width, fill_char)
6
7 print(headline1("type comments work", width=40))
8
9 def headline2(
10     text,          # type: str
11     width=80,      # type: int
12     fill_char='-', # type: str
13 ):              # type: (...) -> str
14     return f" {text.title()} ".center(width, fill_char)
15
16 print(headline2("these type comments also work", width=70))
17 |
```



Typing for sequences, maps, tuples, etc.

```
2   from typing import List, Dict, Tuple  
-  
10  mylist: List[int] = [1, 2, 3, 4]  
11  mydictionary: Dict[str, int] = {'a': 1, 'b': 2, 'c': 3}  
12  coord: Tuple[int, int, int] = (7, 12, 3)  
13  
14  # something fancier  
15  coordinate_list: List[Tuple[int, int]] = []  
16
```

```
mylist.append('a')  
print(mylist)  
  
mydictionary[12] = "hello"  
print(mydictionary)  
  
print(coord)  
coord = (0, 0)  
print(coord)  
  
coordinate_list.append((1,2))  
coordinate_list += [(2,3), (0,0)]  
coordinate_list.append(("hello", "world"))  
print(coordinate_list)
```



Typing for sequences, maps, tuples, etc.

```
(base) [503 uranus:10_patterns1 ] mypy annotations2.py
annotations2.py:17: error: Argument 1 to "append" of "list" has incompatible type "str"; expected "int"
annotations2.py:20: error: Invalid index type "int" for "Dict[str, int]"; expected type "str"
annotations2.py:20: error: Incompatible types in assignment (expression has type "str", target has type "int")
annotations2.py:26: error: Incompatible types in assignment (expression has type "Tuple[int, int]", variable has type "Tuple[int, int, int]")
annotations2.py:31: error: Argument 1 to "append" of "list" has incompatible type "Tuple[str, str]"; expected "Tuple[int, int]"
Found 5 errors in 1 file (checked 1 source file)
```

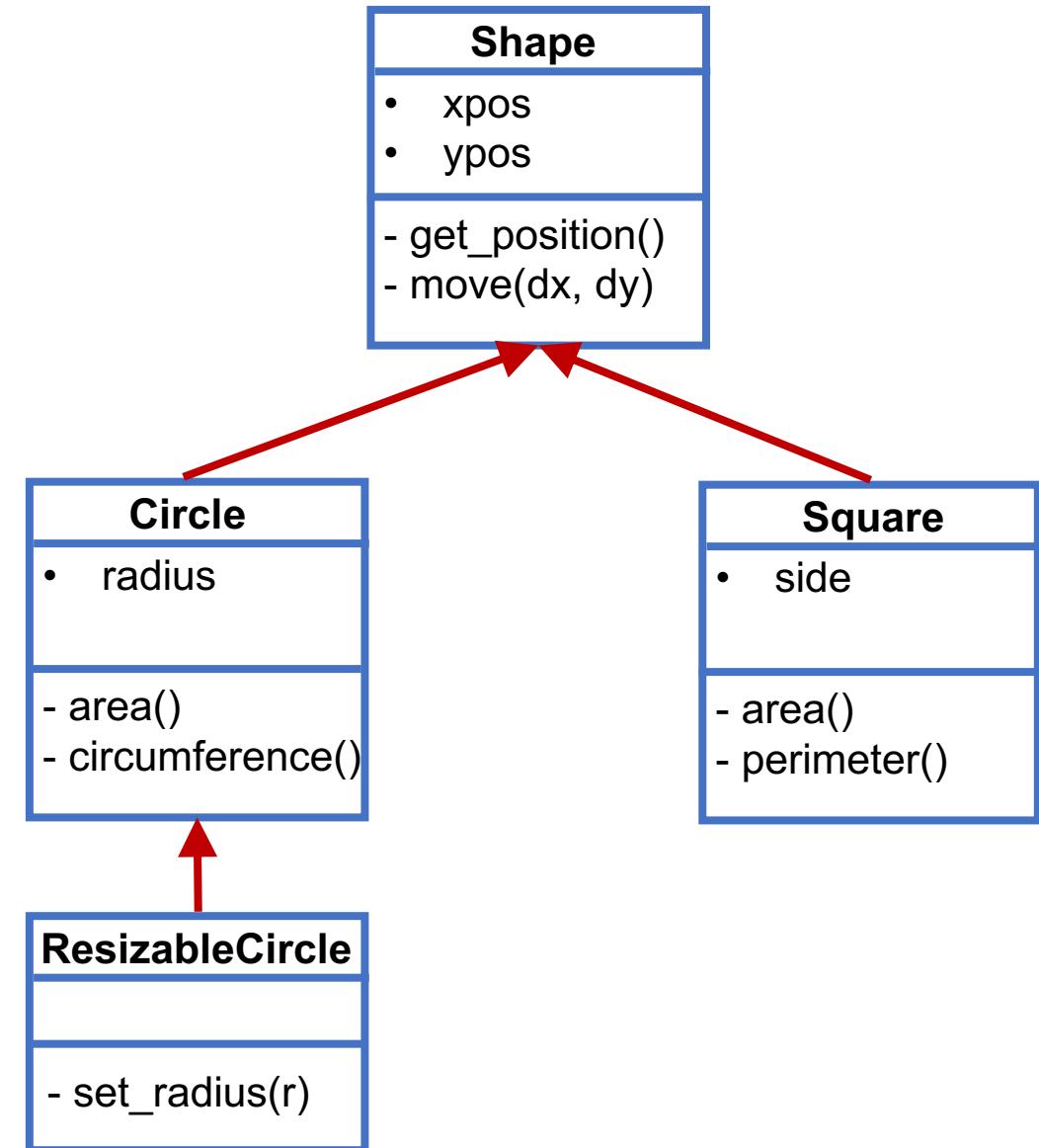


Polymorphism

Polymorphism is the ability to treat a class differently depending on which subclass is being implemented.

For example, if we call the **area** method, the calculation performed depends on which subclass we are using - circle or square.

If we wanted to, a specific subclass of Shape could implement a different way of moving, or it could inherit its behavior from the Shape parent class.



Duck Typing

Duck typing is related to dynamic typing

- **The type or the class of an object is less important than the methods it defines**
- **Instead of checking for the class or type, check the object for the presence of specific methods and/or attributes**



Duck typing: *If it behaves like a duck, it's a duck!*

Python doesn't enforce the type hierarchy when passing parameters. As long as the parameters support the behaviors we need, Python will not complain!

```
def draw_shape(sh):
    """ Draw a shape """
    sh.draw()
    .
    .

draw_shape(not_a_shape) # This works as long
                        # as not_a_shape has
                        # a draw method!
```



Natural Language Processing

Comparative Text Analysis



Northeastern University

Text Analysis Use Cases

Manufacturers

- Identify root causes of product issue quicker
- Identify trends in market segments
- Understand competitors products

Government

- Identify fraud
- Understand public sentiments about unmet needs
- Find emerging concerns that can shape policy

Financial Institutions

- Use contact center transcriptions
- Understand customers
- Identify money laundering or other fraudulent situation

Retail

- Identify profitable customers and understand the reasons for their loyalty
- Manage the brand on social media

Legal

- Identify topics and keywords in discovery documents
- Find patterns in defendant's communications

Healthcare

- Find similar patterns in doctor's reports
- Use social media to detect outbreaks earlier
- Identify patterns in patient claims data

Telecommunications

- Prevent customer churn
- Suggest up-sell/cross-sell opportunities by understanding customer comments

Life Sciences

- Identify adverse events in medicines or vaccines
- Recommend appropriate research materials

Insurance

- Identify fraudulent claims
- Track competitive intelligence
- Manage the brand on social media

zencos 

<https://www.zencos.com/blog/text-mining-examples-advanced-analytics/> <https://www.zencos.com/blog/text-mining-examples-advanced-analytics/>



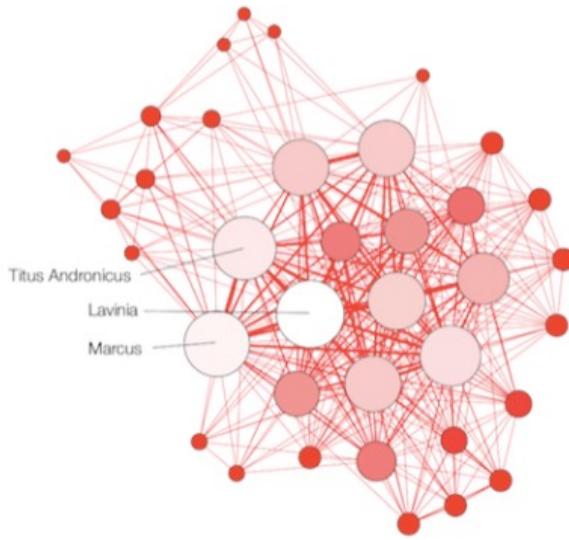
WordClouds



Source: <https://tesseract.academy/a-quick-look-at-the-core-fields-of-data-science/>



Interaction Networks in Shakespeare Plays



TITUS ANDRONICUS

Number of characters **36** | 50% Network density



ROMEO AND JULIET

Number of characters **41** | 37% Network density



JULIUS CAESAR

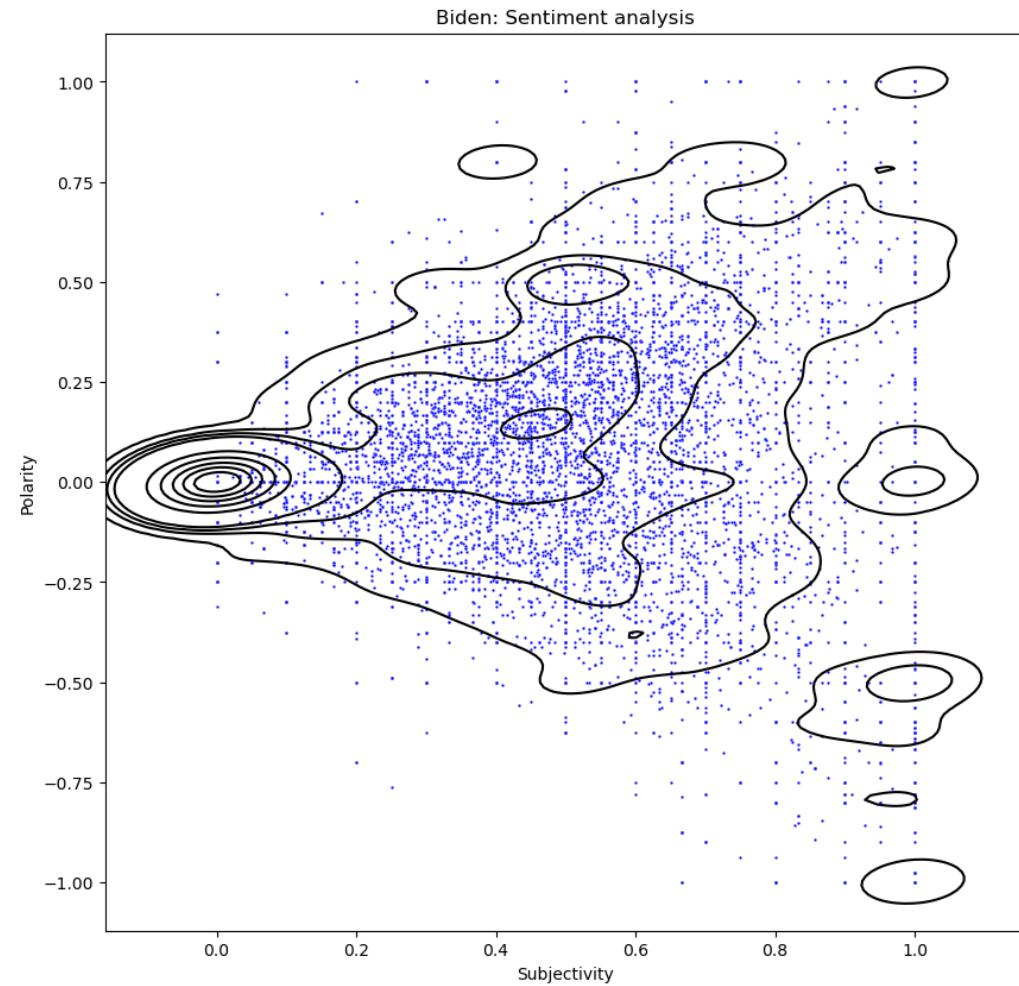
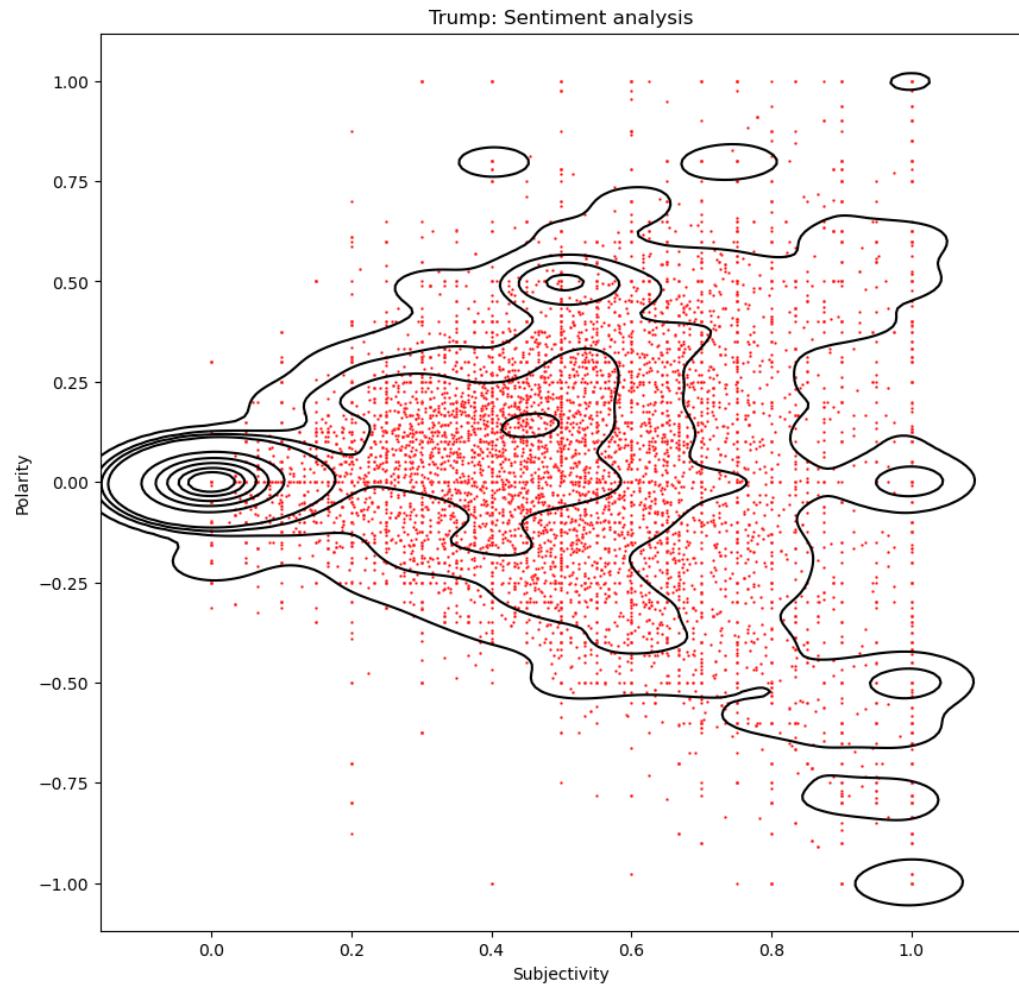
Number of characters **46** | 34% Network density

From: www.martingrandjean.ch/network-visualization-shakespeare



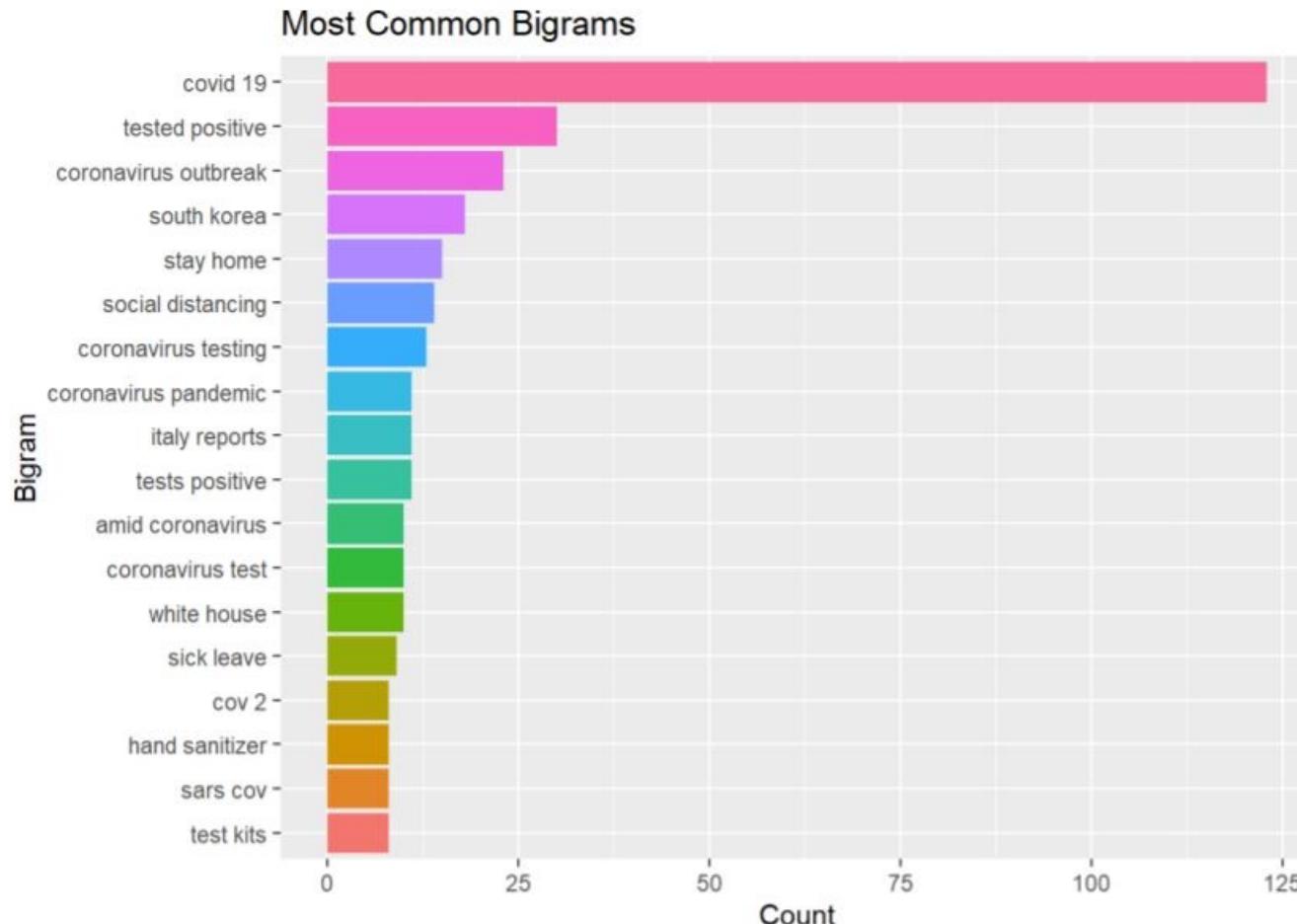
Northeastern University

Political Speech



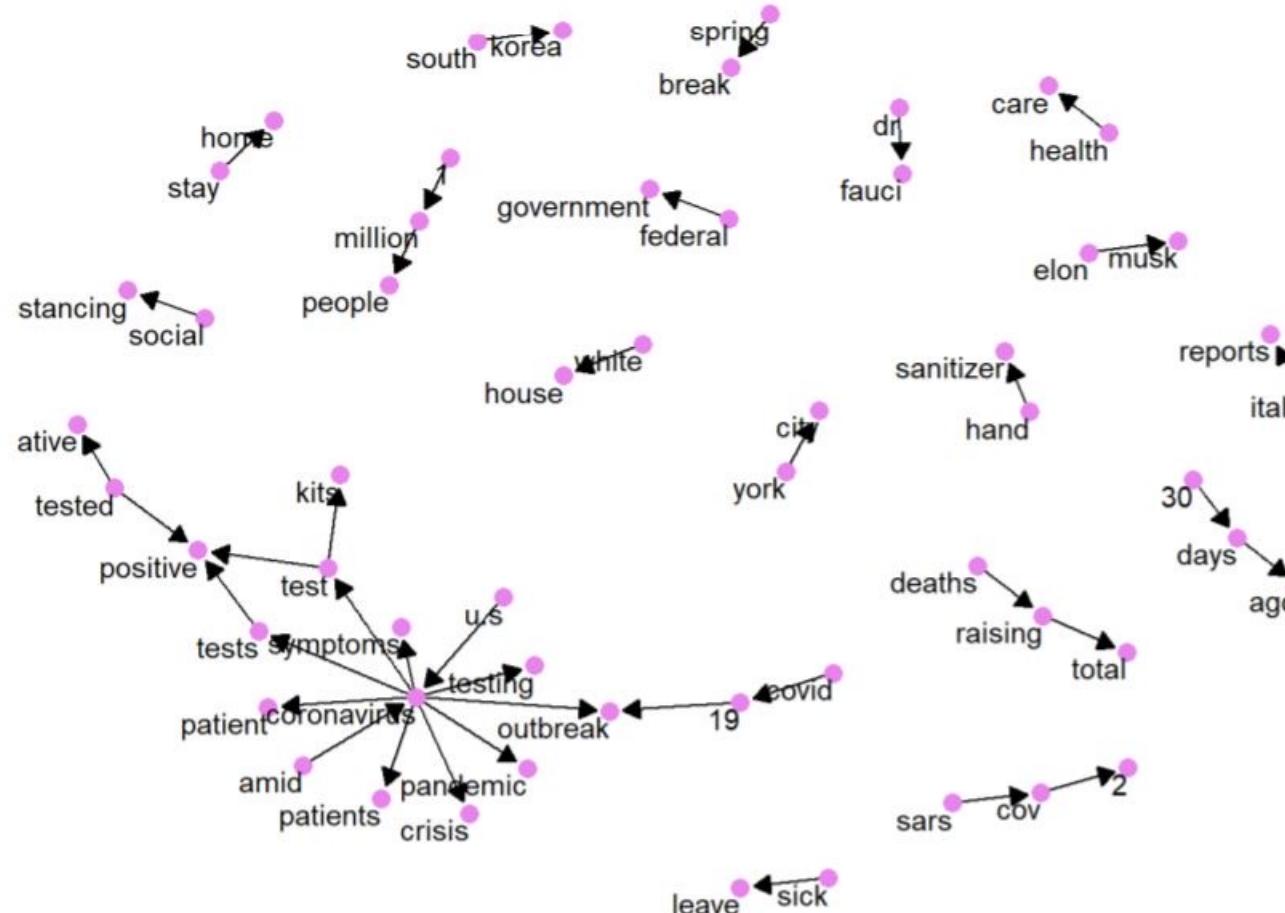
Visualisations – Non-comparative

- Most common bigrams from r/Coronavirus in Mar'20 [1]



Word/Token Networks

- Model bigrams into a network [1]



BiGram Frequency

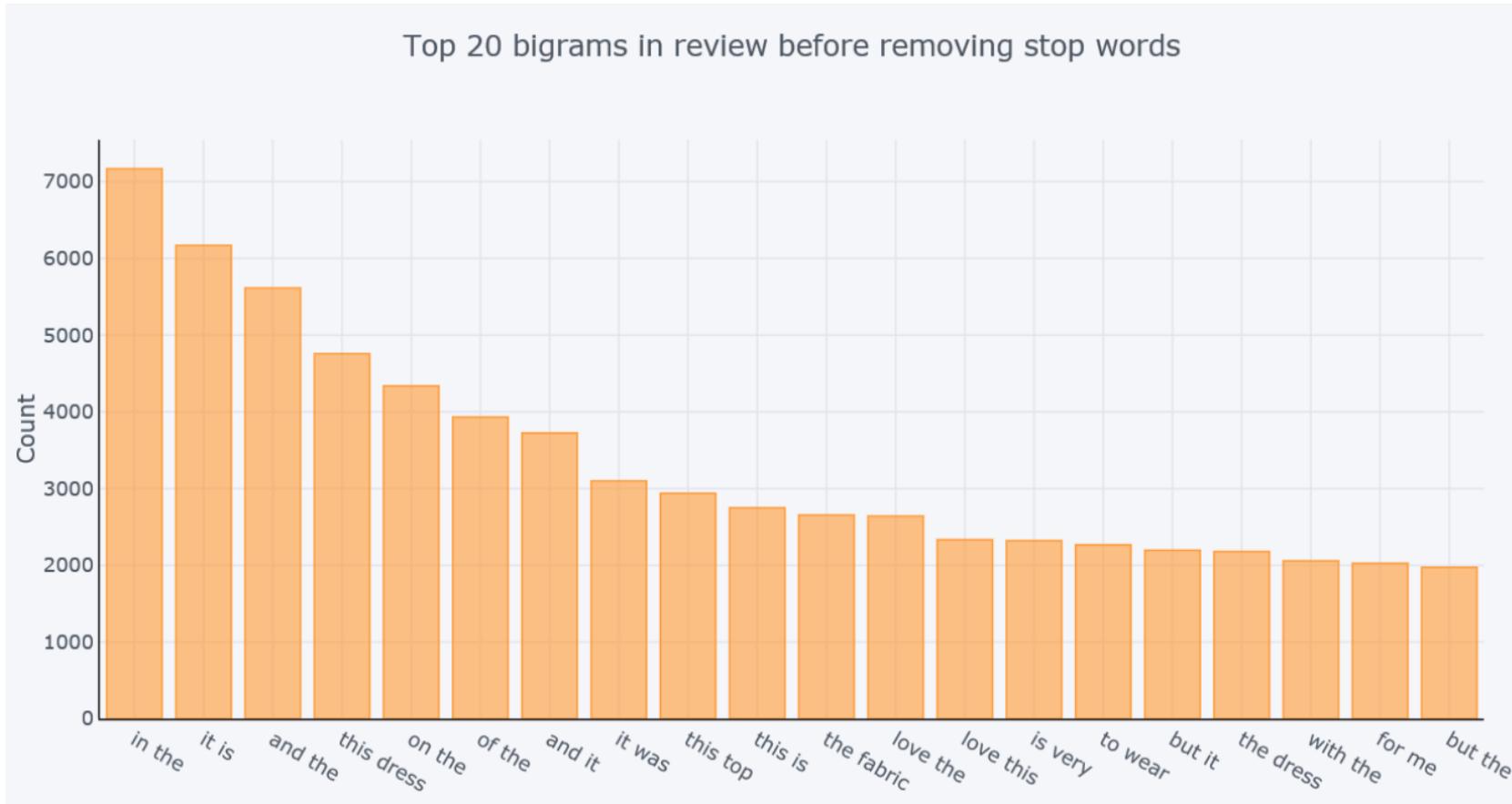


Figure 14

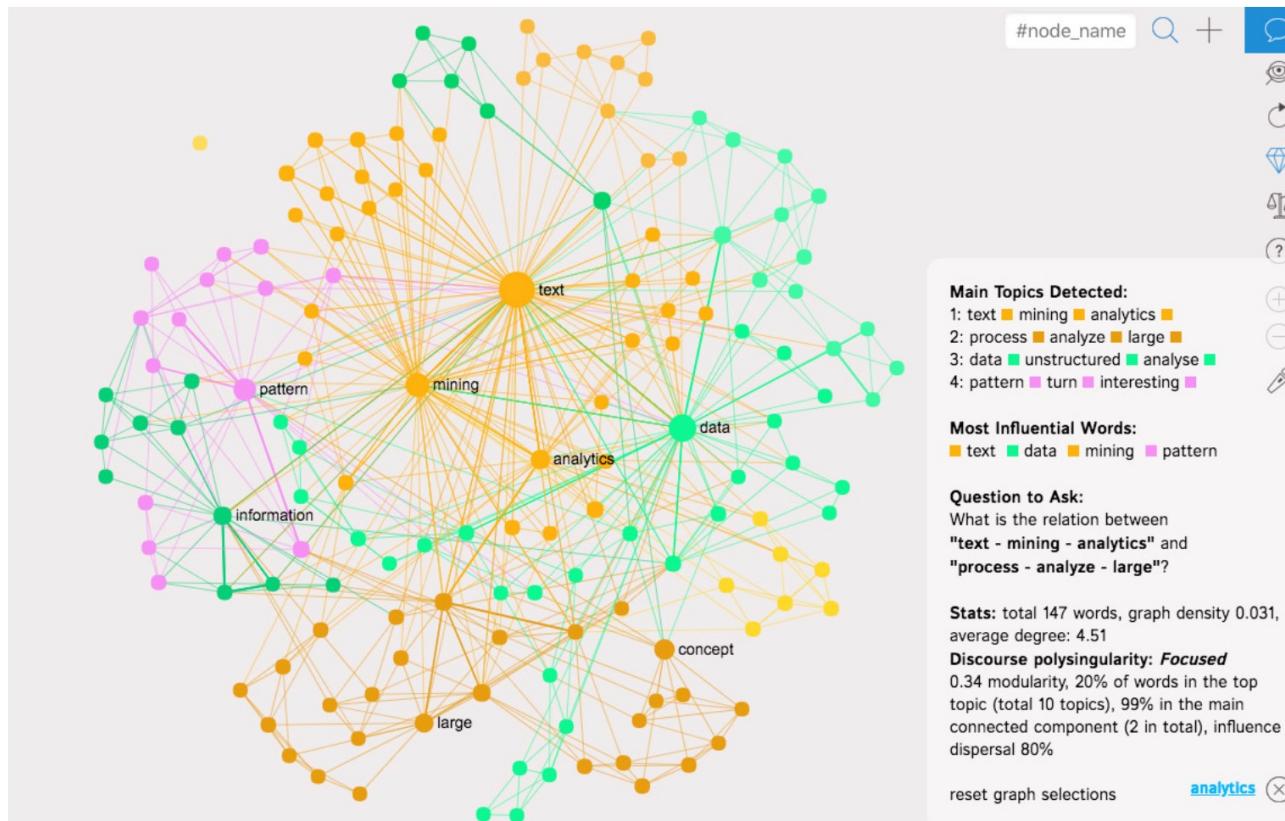
Source: <https://towardsdatascience.com/>



Northeastern University

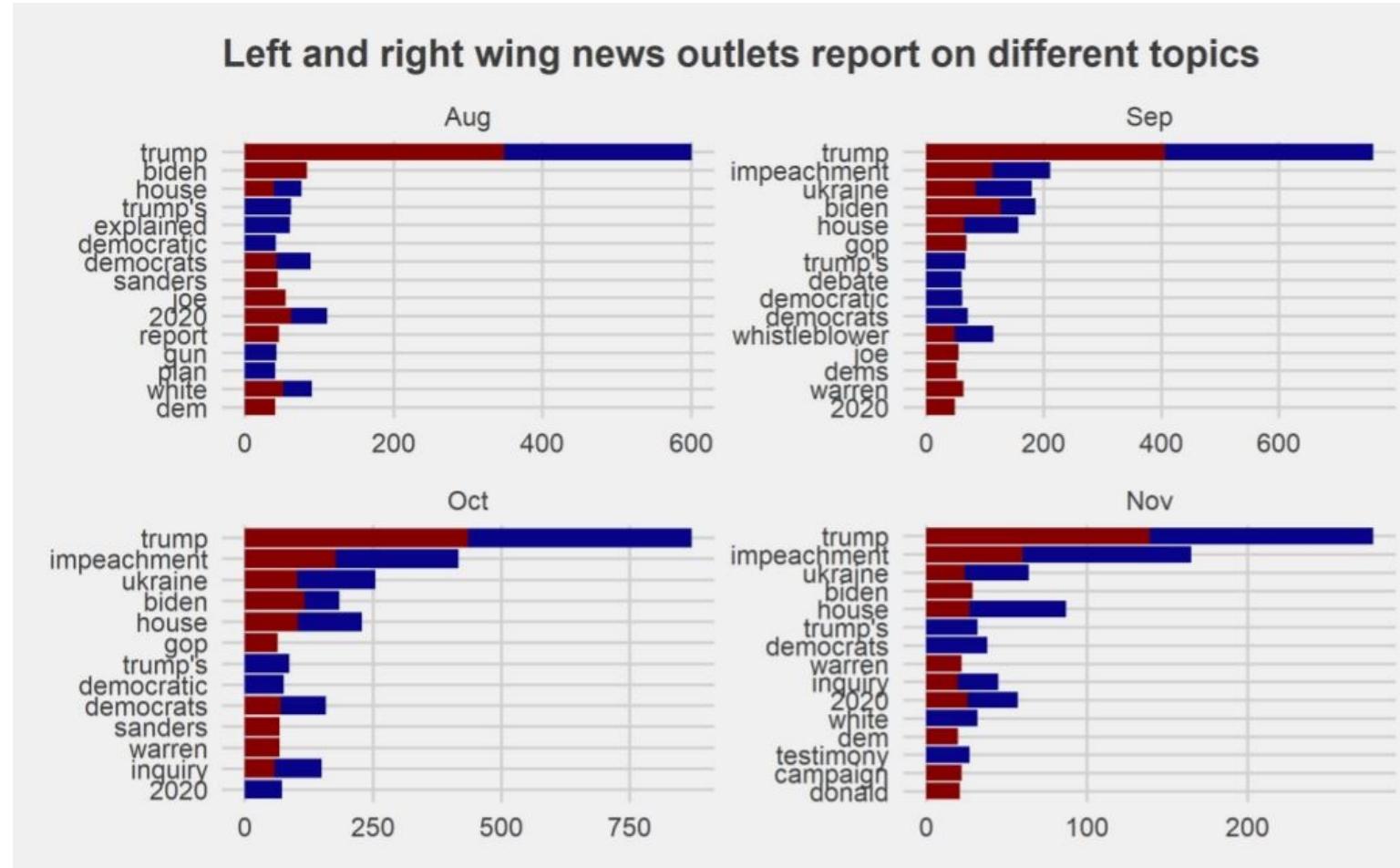
Analyzing Google Search Results

If we use this algorithm to visualize Google search results for the search query “text mining” we get the following network graph visualization:

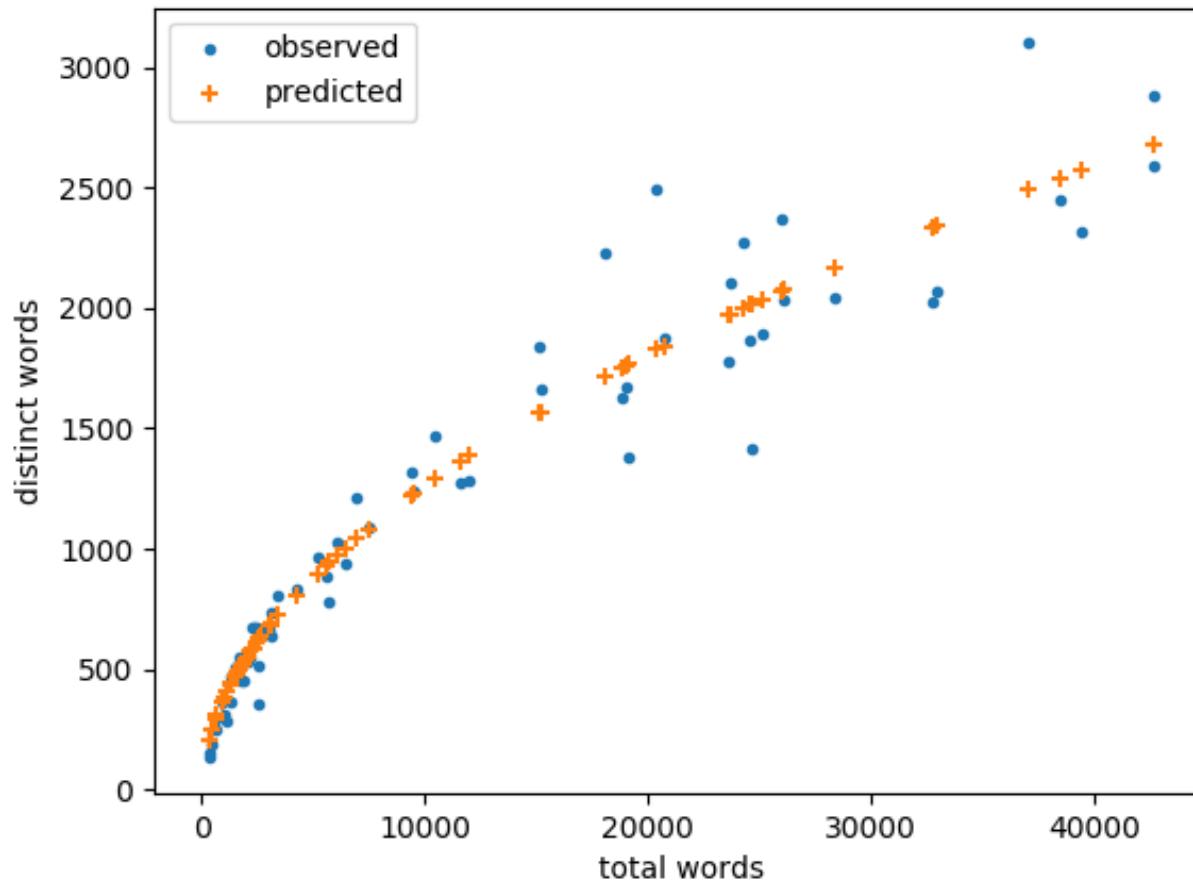


Visualisations – Comparative

- Similarities and differences between news sources in the US [2]



Heaps' Law: $|V| = KN^B$



Zipf's Law

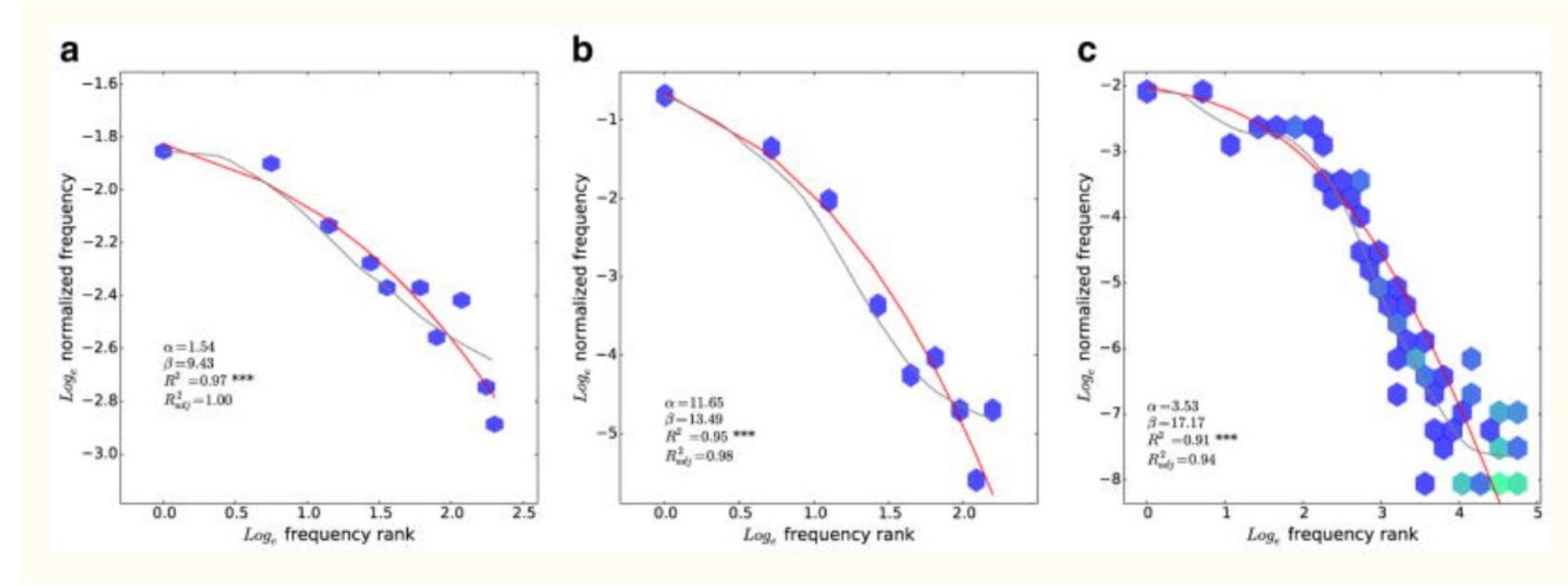
Zipf's law: The r -th most frequent word has a frequency $f(r)$ that scales according to:

$$f(r) \propto \frac{1}{r^\alpha}$$

or more generally:

$$f(r) \propto \frac{1}{(r + \beta)^\alpha}$$

$$\alpha \approx 1 \text{ and } \beta \approx 2.7$$



From: Piantadosi (2015). *Zipf's word frequency law in natural language: A critical review and future directions*. Psychon Bul Rev 21(5): 1112-1130



PROJECT IDEAS



Project Ideas

- Game AI (Chess, Connect4)
- Stock Trading Algorithms
- Pandemic Simulators
- NLP Text Analysis of Song Lyrics
- Bioinformatics: Multi-Sequence Alignment
- Boston Crime Dashboard
- Exploring Gene Function (GO + GeneRif)
- Book Recommendations based on NLP analysis of reviews



Text Analysis Use Cases

Manufacturers

- Identify root causes of product issue quicker
- Identify trends in market segments
- Understand competitors products

Government

- Identify fraud
- Understand public sentiments about unmet needs
- Find emerging concerns that can shape policy

Financial Institutions

- Use contact center transcriptions
- Understand customers
- Identify money laundering or other fraudulent situation

Retail

- Identify profitable customers and understand the reasons for their loyalty
- Manage the brand on social media

Legal

- Identify topics and keywords in discovery documents
- Find patterns in defendant's communications

Healthcare

- Find similar patterns in doctor's reports
- Use social media to detect outbreaks earlier
- Identify patterns in patient claims data

Telecommunications

- Prevent customer churn
- Suggest up-sell/cross-sell opportunities by understanding customer comments

Life Sciences

- Identify adverse events in medicines or vaccines
- Recommend appropriate research materials

Insurance

- Identify fraudulent claims
- Track competitive intelligence
- Manage the brand on social media

zencos 

<https://www.zencos.com/blog/text-mining-examples-advanced-analytics/>

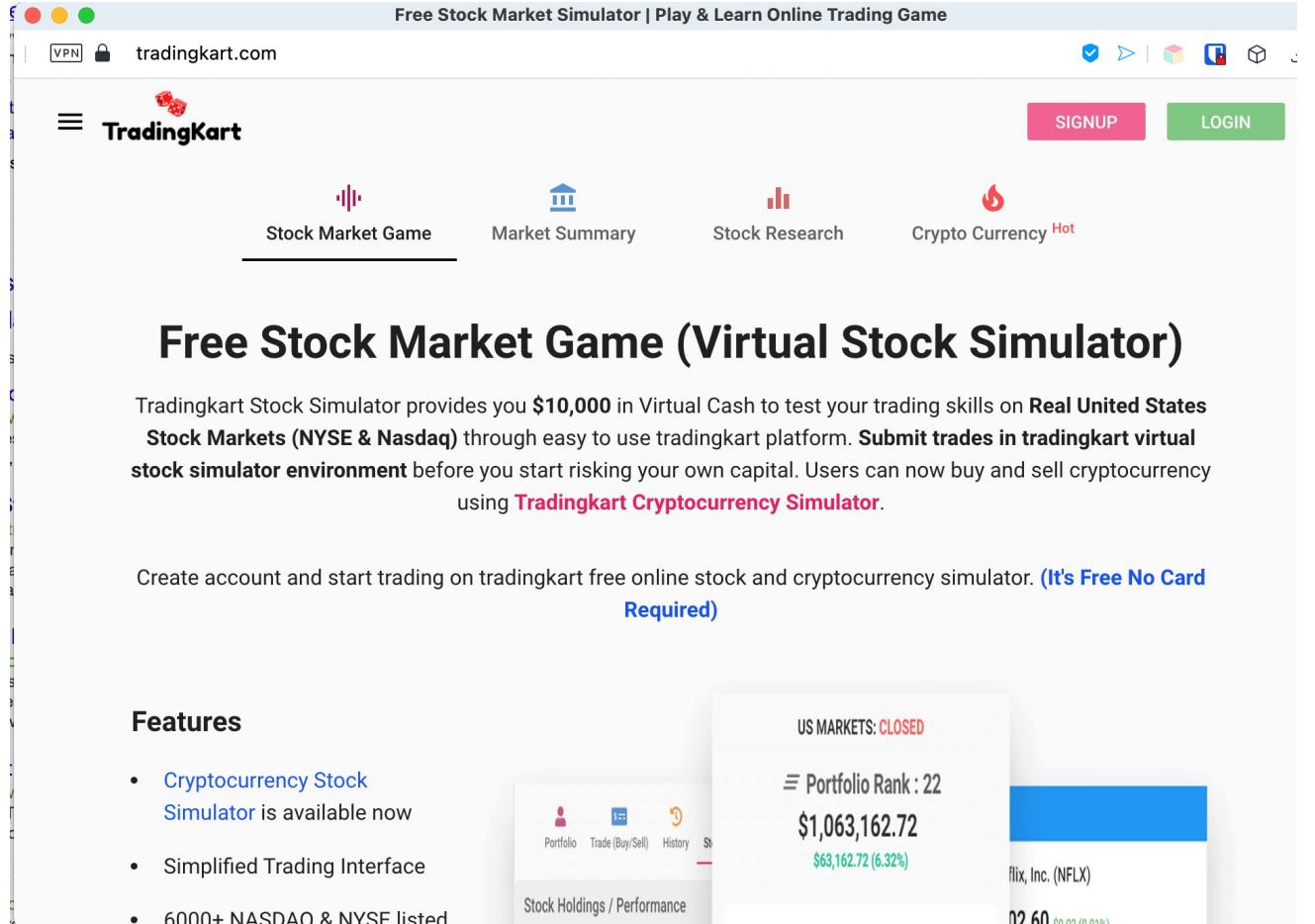
e.g., extend HW3:

- Add interactive dashboards
- Add new visualizations
- In-depth analysis of a larger text collection
- Integrate visualizations from fellow students



Stock market / Trading Bots

- Manage portfolios
- Buy / Sell / Trade
- Develop REST APIs
- Explore trading strategies with bots



The screenshot shows the homepage of TradingKart.com. At the top, there's a navigation bar with a VPN icon, a lock icon, and the URL "tradingkart.com". On the right side of the bar are "SIGNUP" and "LOGIN" buttons. Below the bar, the TradingKart logo is displayed next to a dice icon. The main menu includes "Stock Market Game" (which is underlined), "Market Summary", "Stock Research", and "Crypto Currency" (with a "Hot" tag). To the right of the menu, there are four icons: a bar chart, a building, a bar chart, and a fire icon. A large section titled "Free Stock Market Game (Virtual Stock Simulator)" is centered. It explains that the simulator provides \$10,000 in virtual cash to test trading skills on real US stock markets (NYSE & Nasdaq) through their platform. It encourages users to submit trades in the virtual environment before risking their own capital. It also mentions the "Tradingkart Cryptocurrency Simulator". Below this text, there's a call-to-action: "Create account and start trading on tradingkart free online stock and cryptocurrency simulator. (It's Free No Card Required)". On the right side of the page, there's a sidebar with a "US MARKETS: CLOSED" message, a "Portfolio Rank: 22", and a "Stock Holdings / Performance" section showing a portfolio value of \$1,063,162.72, a gain of \$63,162.72 (6.32%), and a recent trade for Netflix (NFLX) at 02.60 with a gain of \$0.03 (0.01%).



Sports Analytics Dashboards

The Warriors and Curry

With the Warriors setting a NBA record for most consecutive wins to start a season it was time to have a closer look at Steph Curry.

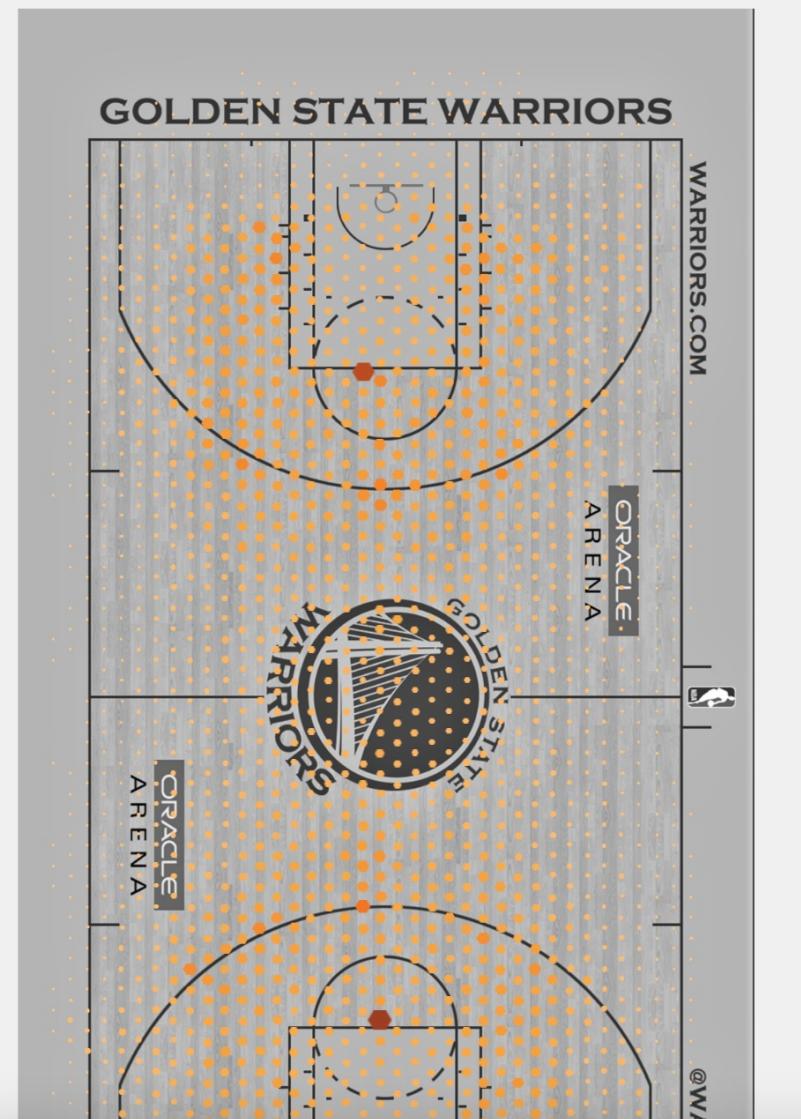
Select the game you want to analyse and see how Steph Curry moved round the court along with the points, rebounds, steals and assists he contributed to the win.



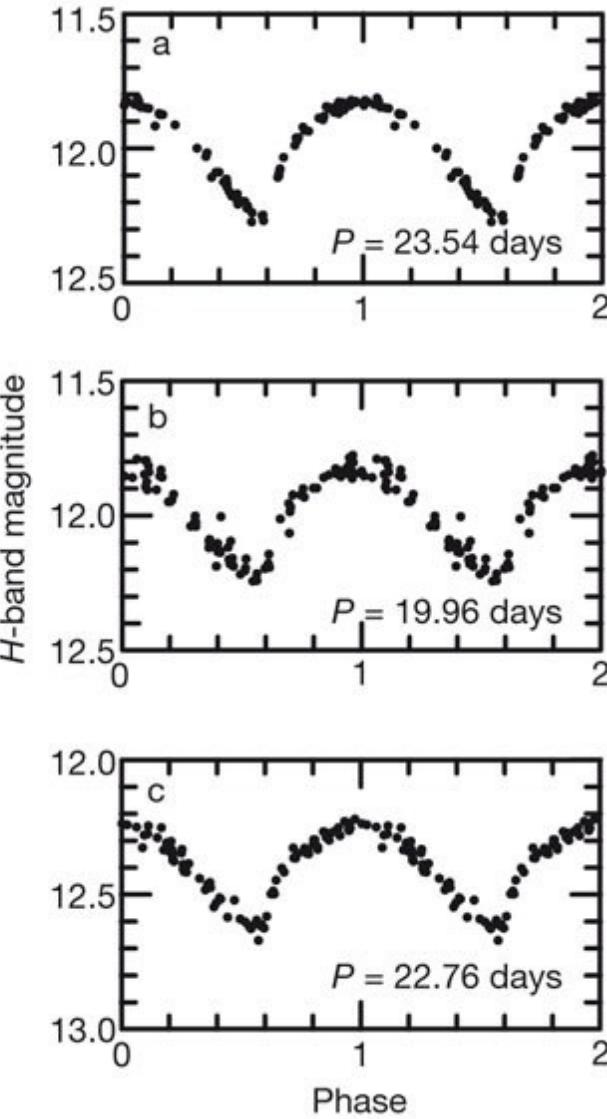
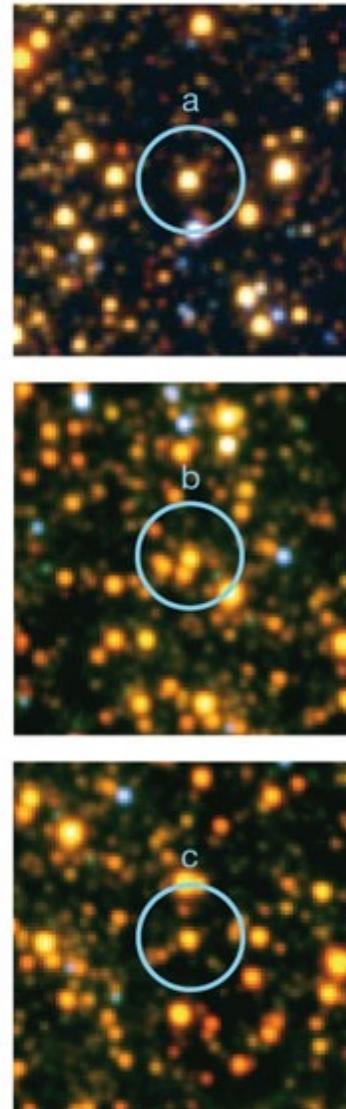
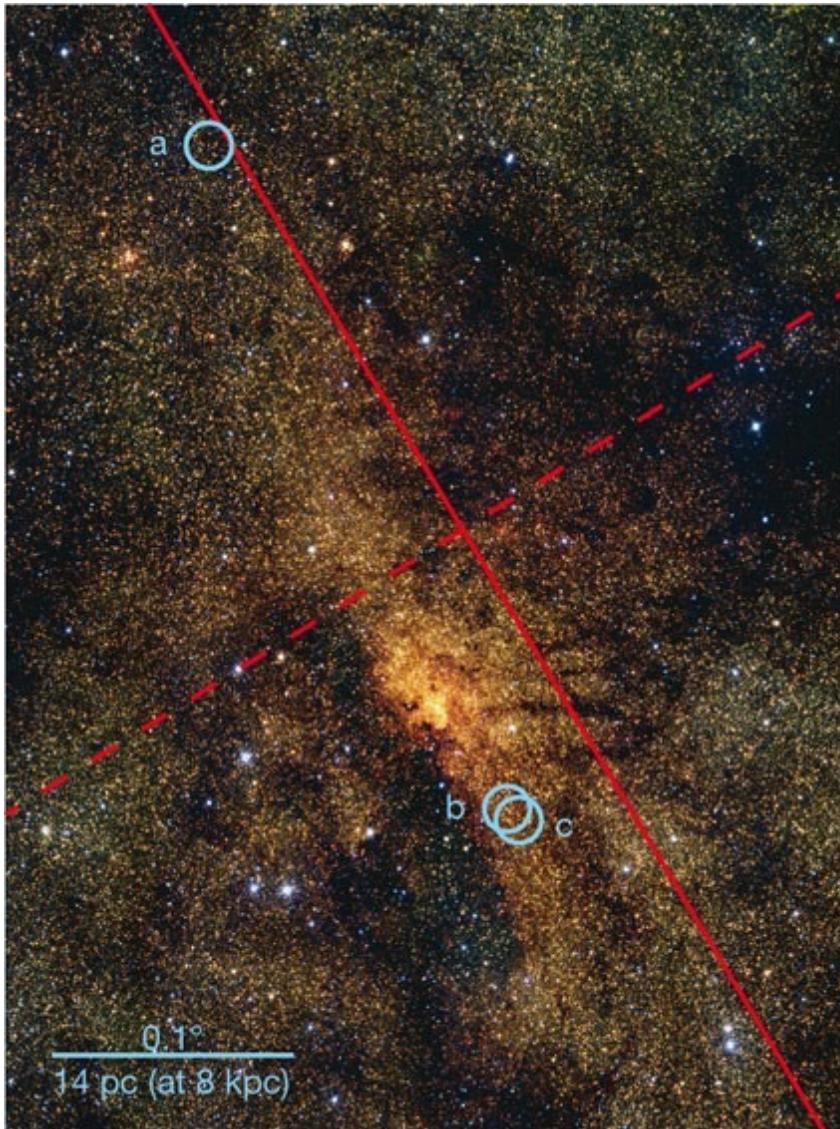
Select who Curry tortured or [click](#) on the bar chart below to filter everything else

▼

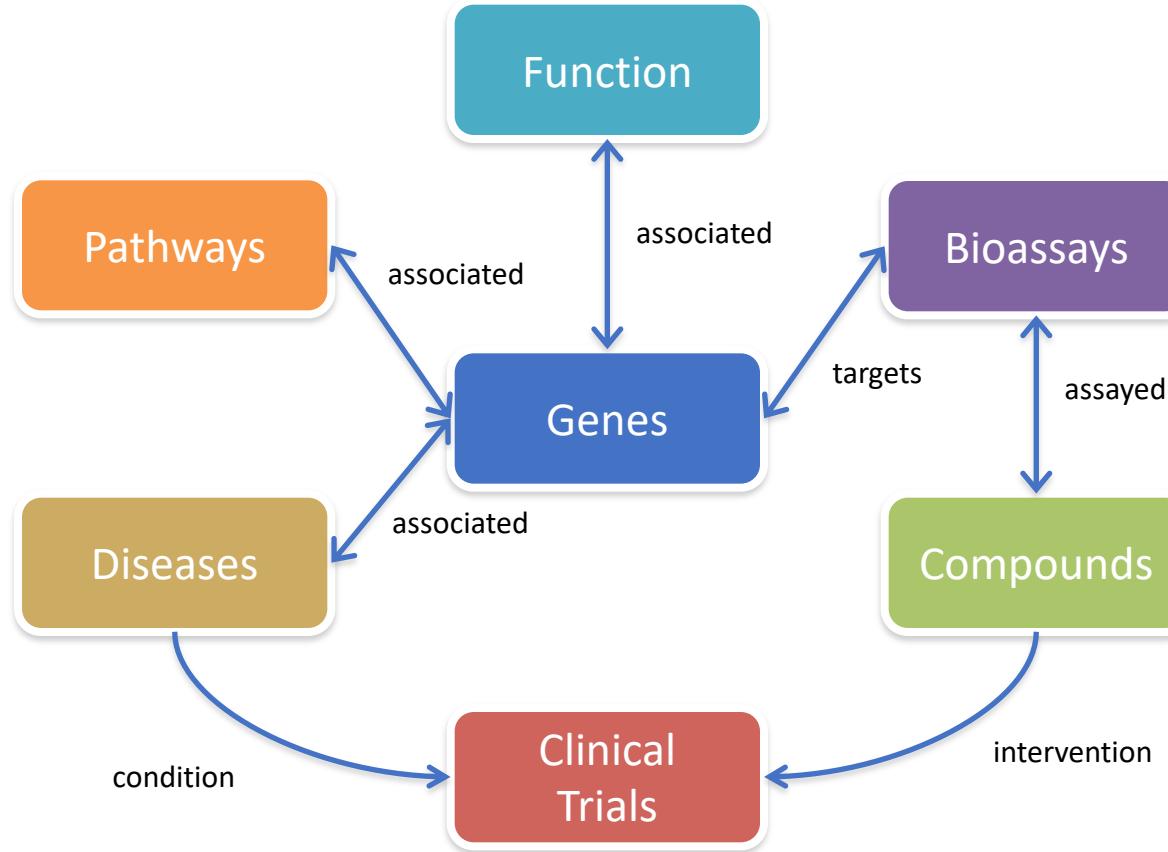
Points per Game vs his own Season Average



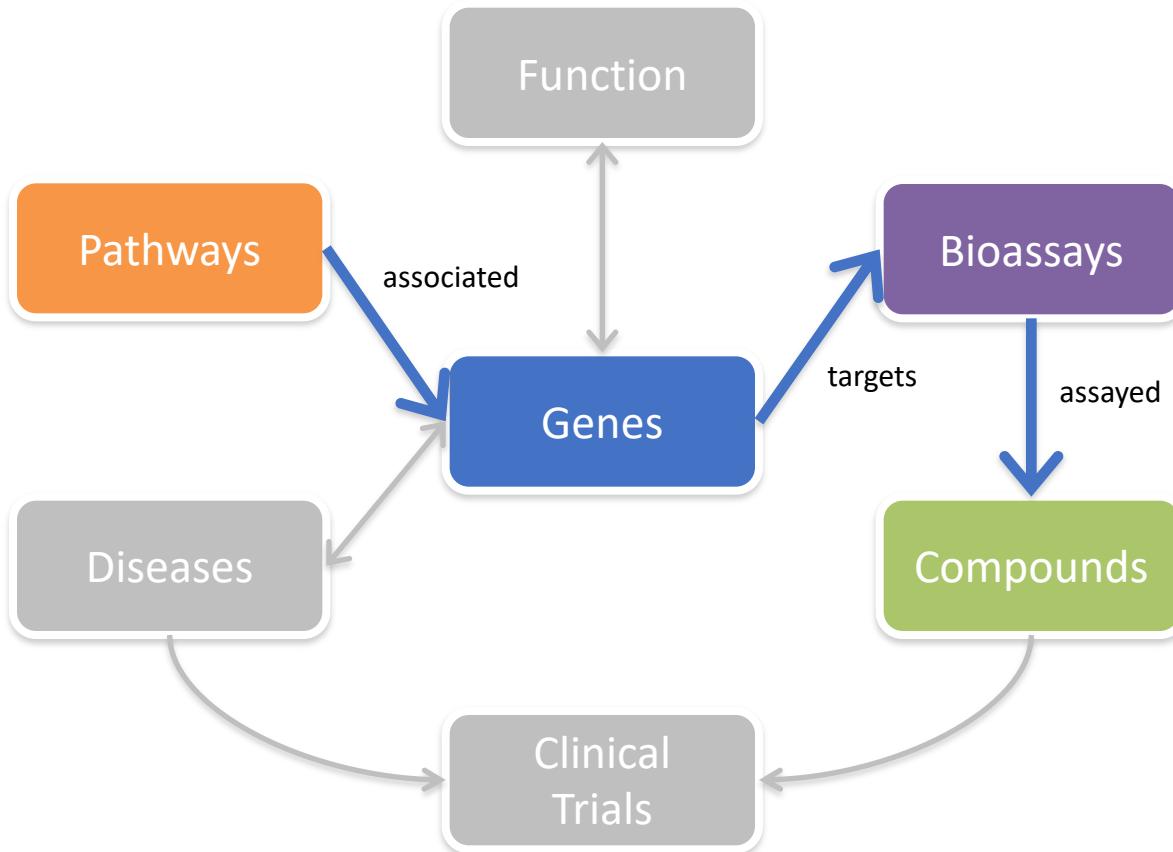
Variable Stars Database and Dashboard



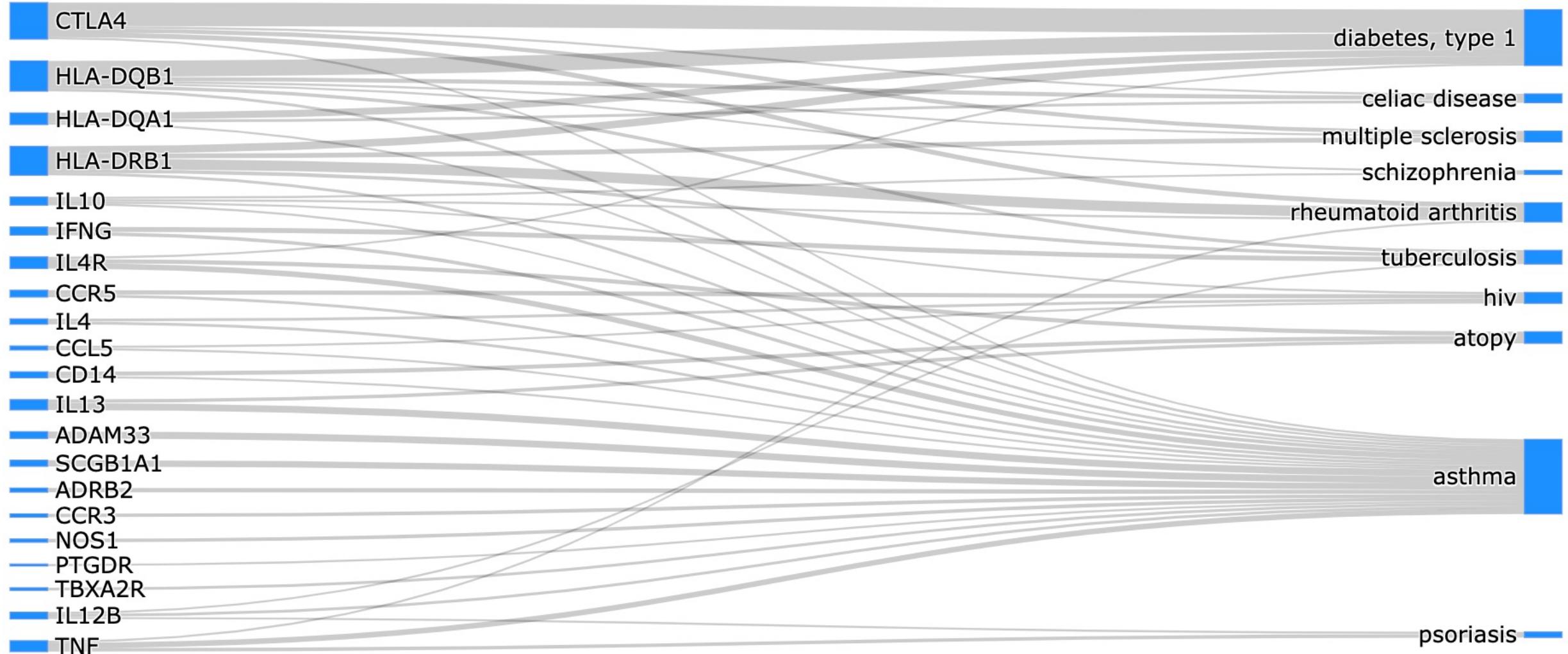
Biological Knowledge is Connected



Query: What compounds activate genes association with pathway X?



Biolink



The BioLink DataTable Widget

A configurable container of data

- Title
- Description (help)
- Default Sorting
- Selection ON/OFF
- Actions
- Column order
- Column headers
- URL configuration
- Text wrapping
- Export ON/OFF

Associated Genes (25)				Actions ▾
	EntrezGene ID ▲	Gene Symbol ▲	Gene Name	
<input type="checkbox"/>	329	BIRC2	baculoviral IAP repeat containing 2	
<input checked="" type="checkbox"/>	330	BIRC3	baculoviral IAP repeat containing 3	
<input type="checkbox"/>	841	CASP8	caspase 8, apoptosis-related cysteine peptidase	
<input checked="" type="checkbox"/>	857	CAV1	caveolin 1, caveolae protein, 22kDa	
<input checked="" type="checkbox"/>	1540	CYLD	cylindromatosis (turban tumor syndrome)	
<input type="checkbox"/>	4214	MAP3K1	mitogen-activated protein kinase kinase kinase 1	
<input type="checkbox"/>	4215	MAP3K3	mitogen-activated protein kinase kinase kinase 3	
<input checked="" type="checkbox"/>	4217	MAP3K5	mitogen-activated protein kinase kinase kinase 5	
<input type="checkbox"/>	5606	MAP2K3	mitogen-activated protein kinase kinase 3	
<input type="checkbox"/>	5609	MAP2K7	mitogen-activated protein kinase kinase 7	
<input type="checkbox"/>	6609	SMPD1	sphingomyelin phosphodiesterase 1, acid lysosomal	
<input type="checkbox"/>	6610	SMPD2	sphingomyelin phosphodiesterase 2, neutral membrane (neutral sphingomyelinase)	
<input checked="" type="checkbox"/>	6772	STAT1	signal transducer and activator of transcription 1, 91kDa	
<input checked="" type="checkbox"/>	6868	ADAM17	ADAM metallopeptidase domain 17	
<input type="checkbox"/>	7128	TNFAIP3	tumor necrosis factor, alpha-induced protein 3	
<input type="checkbox"/>	7186	TRAF2	TNF receptor-associated factor 2	
<input type="checkbox"/>	7245	TYN	thioredoxin	



An Example Biolink Search

Search: casp*

Working Set: 0 genes (clear) | Guest ▾

BIO LINK EXPLORER

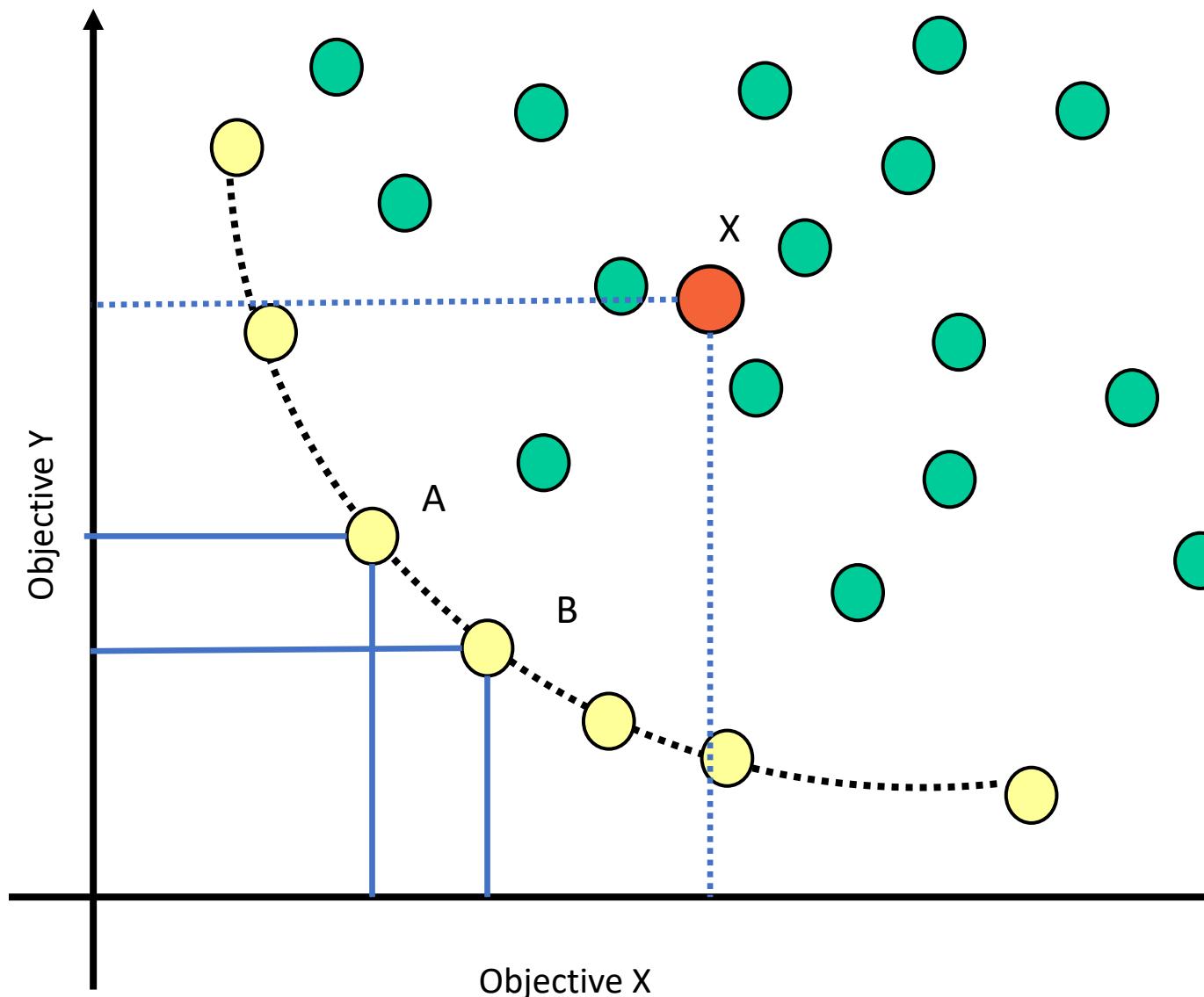
Search: casp*

Examples: mapk, gpcr*, "TNF", breast cancer, HIV, *signaling

Gene (55)						Function (4)	Pathway (1)	Disease (0)	Bioassay (8)	References (441)
Gene Search Results (55)										
	EntrezGene ID	Gene Symbol	Synonyms	Name	Other Names	Actions ▾				
<input type="checkbox"/>	8915	BCL10	CARMEN, CIPER, CLAP, c-E10, mE10	B-cell CLL/lymphoma 10	B-cell lymphoma/leukemia 10, CARD containing molecule enhancing NF- κ B, CARD-containing apoptotic signaling protein, CARD-containing molecule enhancing NF- κ B, CARD-containing proapoptotic protein, CED-3/ICH-1 prodomain homologous E10-like regulator, OTTHUOMP00000011647, OTTHUOMP00000036080, bcl-10, cCARMEN, caspase-recruiting domain-containing protein, cellular homolog of vCARMEN, cellular-E10, hCLAP, mammalian CARD-containing adapter molecule E10					
<input type="checkbox"/>	29775	CARD10	BIMP1, CARMA3, MGC142219	caspase recruitment domain family, member 10	Bcl10 binding protein and activator of NF κ B, CARD-containing MAGUK 3 protein, CARD-containing MAGUK protein 3, OTTHUOMP00000197888, carma 3, caspase recruitment domain-containing protein 10					
<input type="checkbox"/>	84433	CARD11	BIMP3, CARMA1, MGC133069	caspase recruitment domain family, member 11	CARD-containing MAGUK protein 1, bcl10-interacting maguk protein 3, card-maguk protein 1, carma 1, caspase recruitment domain-containing protein 11					
<input type="checkbox"/>	79092	CARD14	BIMP2, CARMA2	caspase recruitment domain family, member 14	CARD-containing MAGUK 2 protein, CARD-containing MAGUK protein 2, bcl10-interacting maguk protein 2, card-maguk protein 2, carma 2, caspase recruitment domain-containing protein 14					
<input type="checkbox"/>	114769	CARD16	COP, COP1, PSEUDO-ICE	caspase recruitment domain family, member 16	CARD only domain-containing protein 1, CARD only protein, caspase recruitment domain-containing protein 16, caspase-1 dominant-negative inhibitor pseudo-ICE, caspase-1 inhibitor COP, pseudo interleukin-1 beta converting enzyme, pseudo interleukin-1beta converting enzyme					
<input type="checkbox"/>	440068	CARD17	INCA	caspase recruitment domain family, member 17	Inhibitory CARD, caspase recruitment domain-containing protein 17, caspase-1 inhibitor INCA, inhibitory caspase recruitment domain (CARD) protein, inhibitory caspase recruitment domain protein					
<input type="checkbox"/>	59082	CARD18	ICEBERG, UNQ5804, pseudo-ICE	caspase recruitment domain family, member 18	ICEBERG caspase-1 inhibitor, caspase recruitment domain-containing protein 18, caspase-1 inhibitor Iceberg					
<input type="checkbox"/>	84674	CARD6	CINCIN1	caspase recruitment domain family, member 6	CARD-containing inhibitor of Nod1 and Cardiak-induced NF- κ B activation, caspase recruitment domain protein 6, caspase recruitment domain-containing protein 6					
<input type="checkbox"/>	22900	CARD8	CARDINAL, DACAR, DAKAR, DKFZp779L0366, FLJ18119, FLJ18121, KIAA0955, MGC57162, NDPP, NDPP1, TUCAN	caspase recruitment domain family, member 8	CARD inhibitor of NF- κ B-activating ligands, apoptotic protein NDPP1, caspase recruitment domain-containing protein 8, tumor up-regulated CARD-containing antagonist of CASP9, tumor up-regulated CARD-containing antagonist of caspase nine					



Solving Hard Multi-Objective Optimization Problems



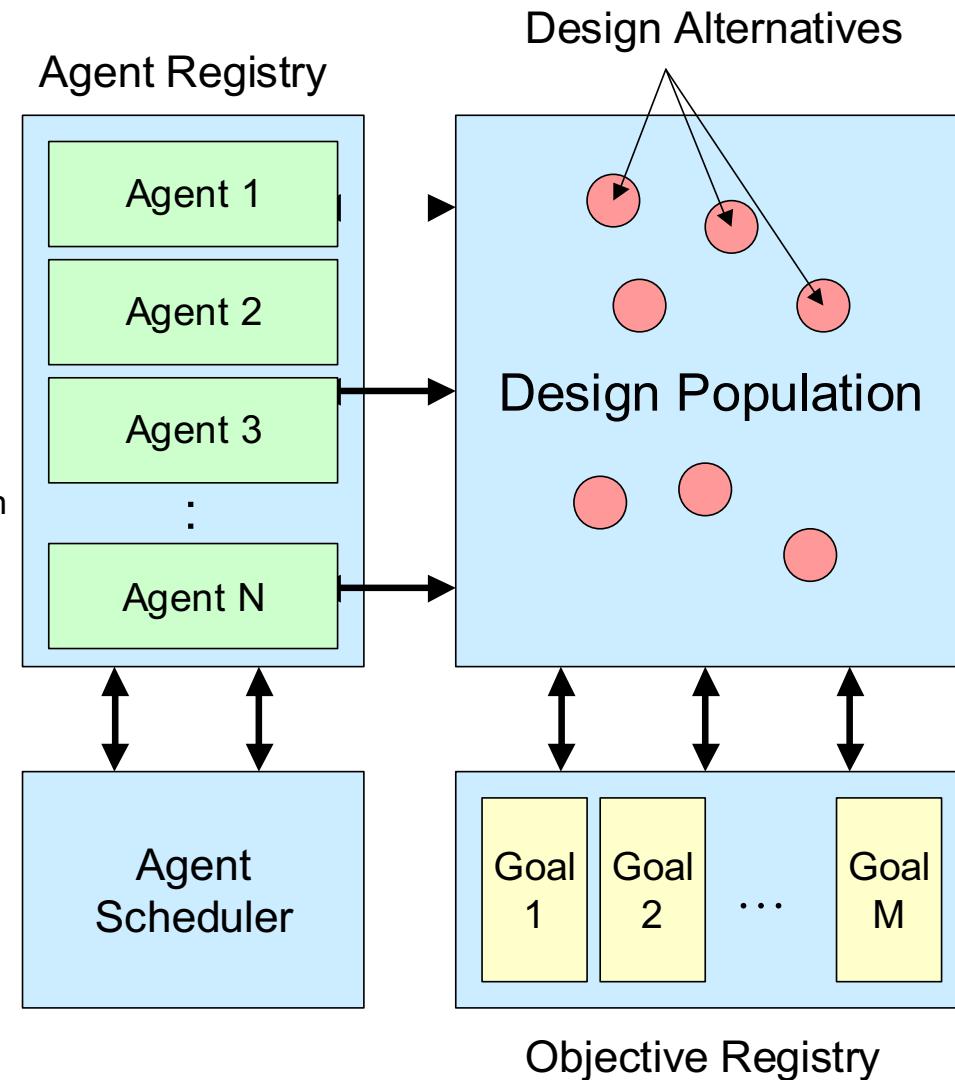
An architecture for multi-objective optimization using evolutionary computing

Classes of agents:

Creator Agent: Adds a design constructed from scratch

Improver Agent: Copies and then modifies an existing design

Destroyer Agent: Removes one or more unpromising designs



General Distance Measure

$$\text{diff} = w_1 |\Delta A_1|^r + w_2 |\Delta A_2|^r + \dots + w_n |\Delta A_n|^r$$

where

ΔA_i is the difference with respect to feature i,

w_k is a feature weighting factor (0.0 to 1.0), and

r is an exponent (> 0.0)



Tweaking / modifying a solution

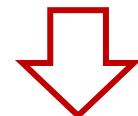
weights

r

k

1.0	1.0	1.0	1.0	2	1
-----	-----	-----	-----	---	---

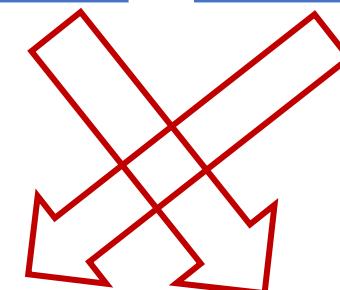
Mutate:



1.0	1.0	0.95	1.0	2	1
-----	-----	------	-----	---	---

0.20	0.90	1.00	0.45	2	5	0.38	0.29	1.00	0.93	1.5	3
------	------	------	------	---	---	------	------	------	------	-----	---

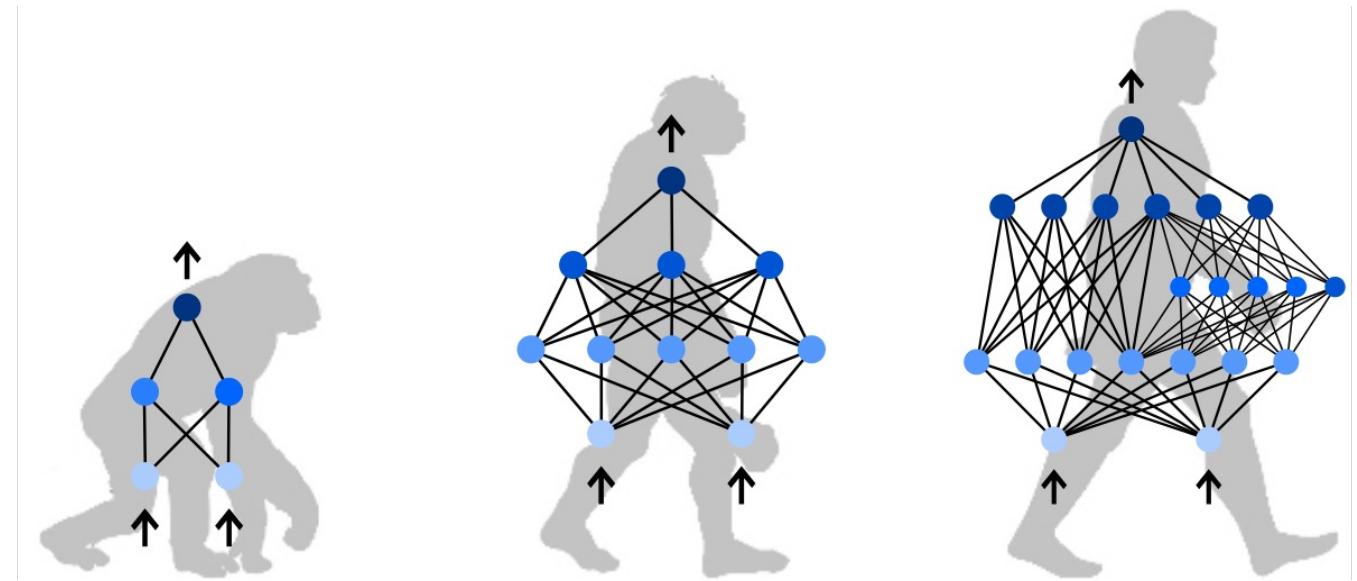
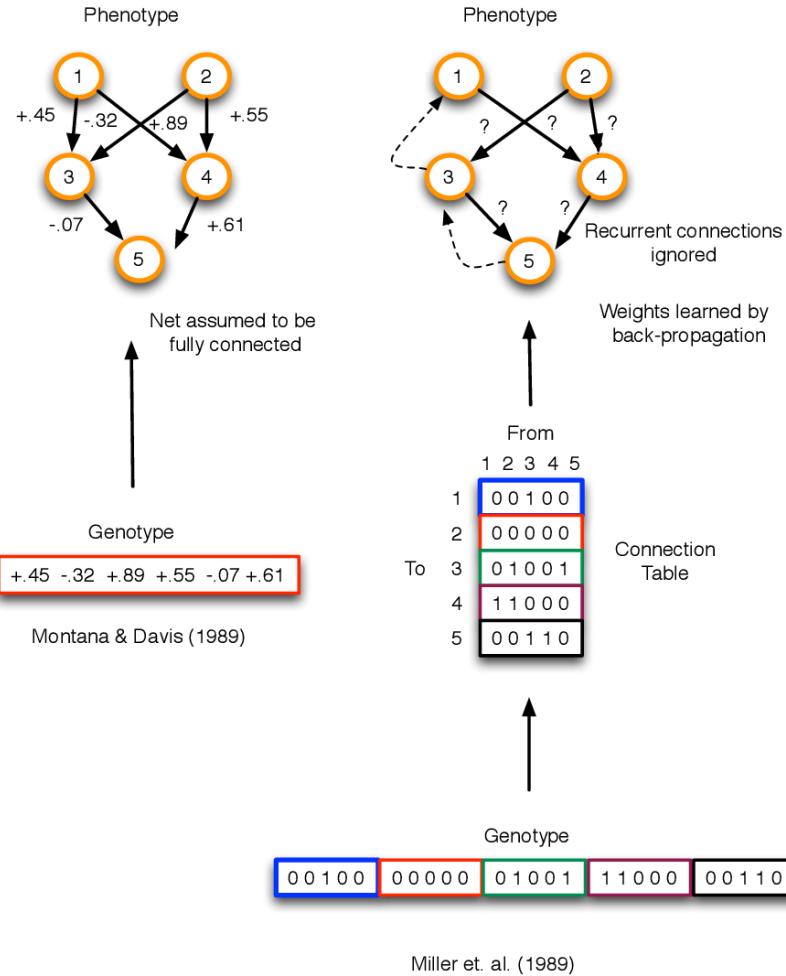
Crossover:



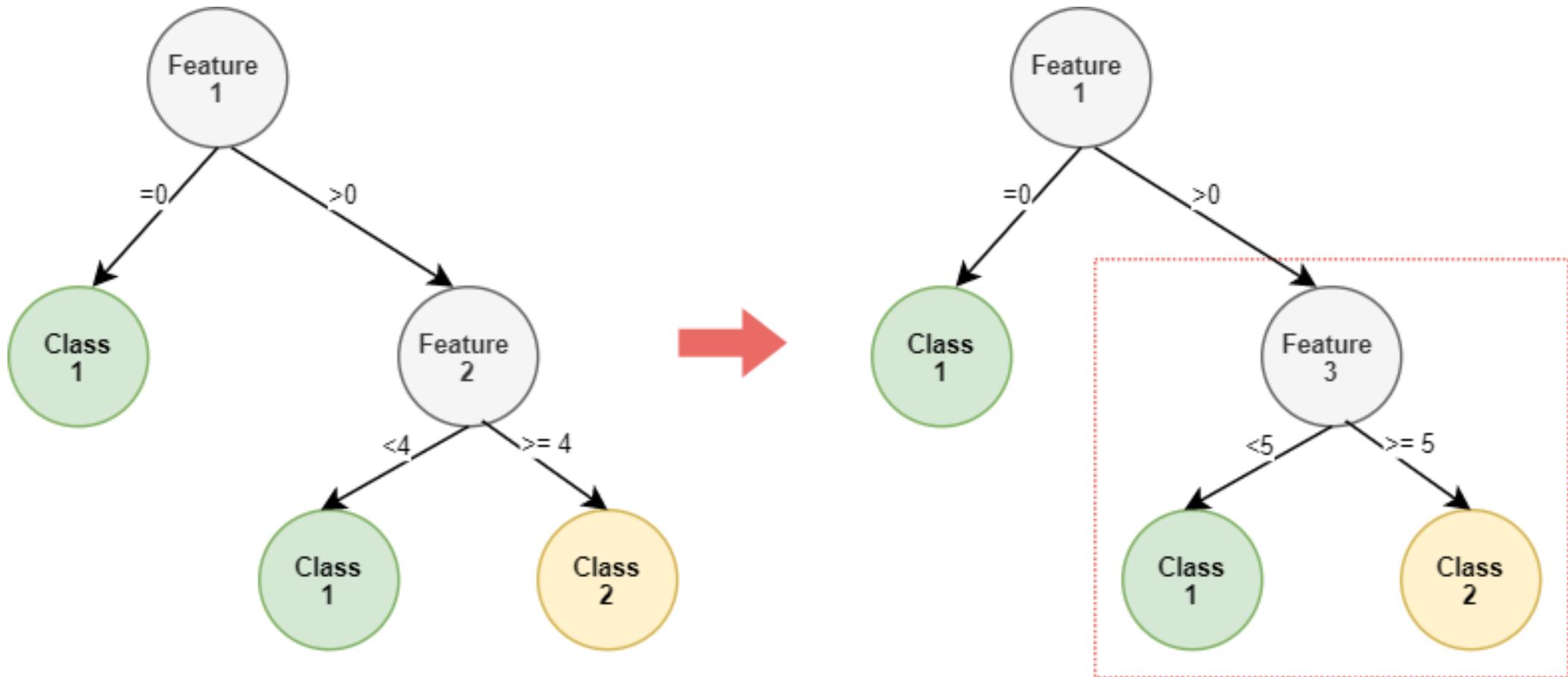
0.20	0.90	1.00	0.93	1.5	3	0.38	0.29	1.00	0.45	2	5
------	------	------	------	-----	---	------	------	------	------	---	---



Evolving Neural Networks

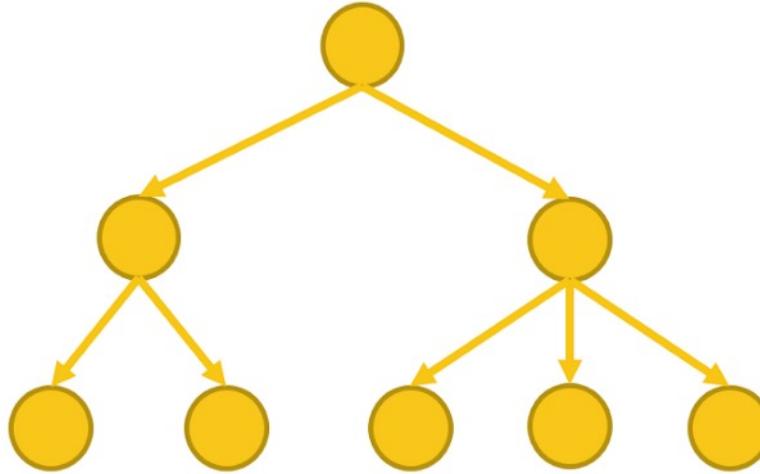


Optimizing Decision Trees

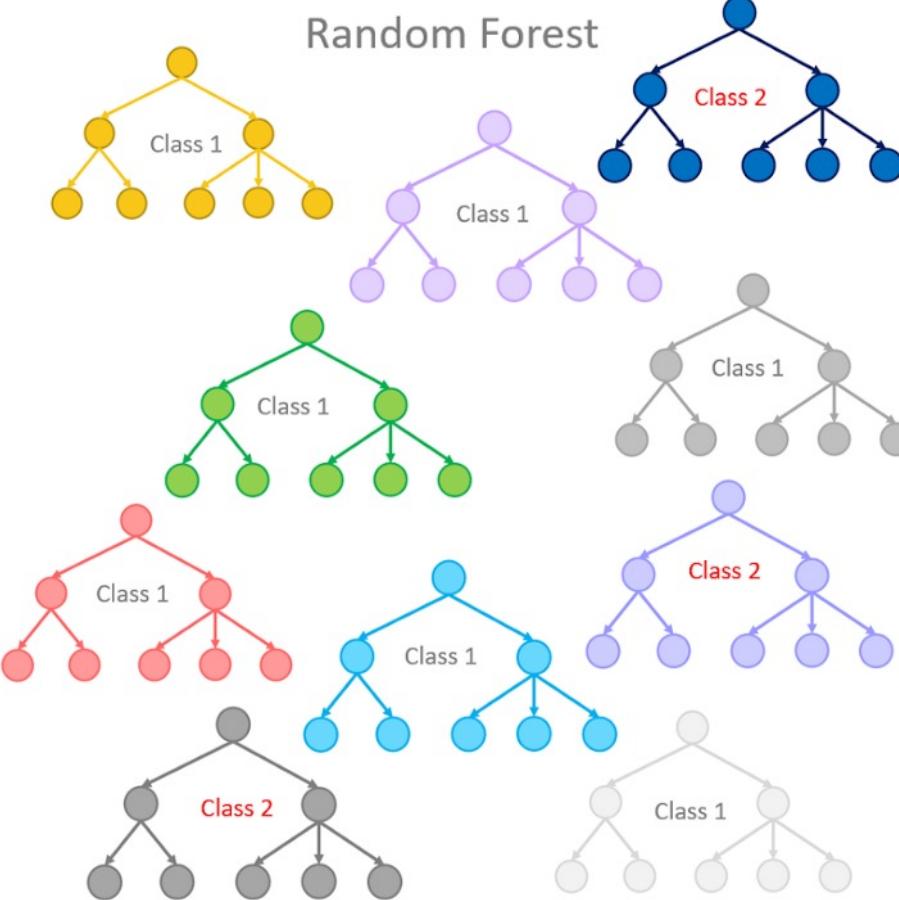


Random Forests

Single Decision Tree



Random Forest



Functional Programming

Elegant Code in Python



Northeastern University

An overview of functional programming

- What is functional programming?
- Immutable data structures
- Building blocks: *filter()*, *map()*, *reduce()*
- Parallel processing, multi-processing, futures
- Emphasis on recursion rather than looping
- list comprehensions everywhere!
- Functions are first-class objects: Can be input to and output from other functions or be elements of a collection.

A large black lambda symbol (λ) centered within a white square frame.

What is functional programming?

Functional programming is a programming *paradigm* or programming *style* that emphasizes computation on *immutable (unchanging)* data structures through the evaluation of functions while avoiding side-effects.

Advantages:

- Reduced number of bugs
- More maintainable
- Elegant expression of computation
- Easier to reason about multi-threaded / parallel programs because interactions with data are via immutable data structures.



Clojure



Haskell



Erlang



Racket

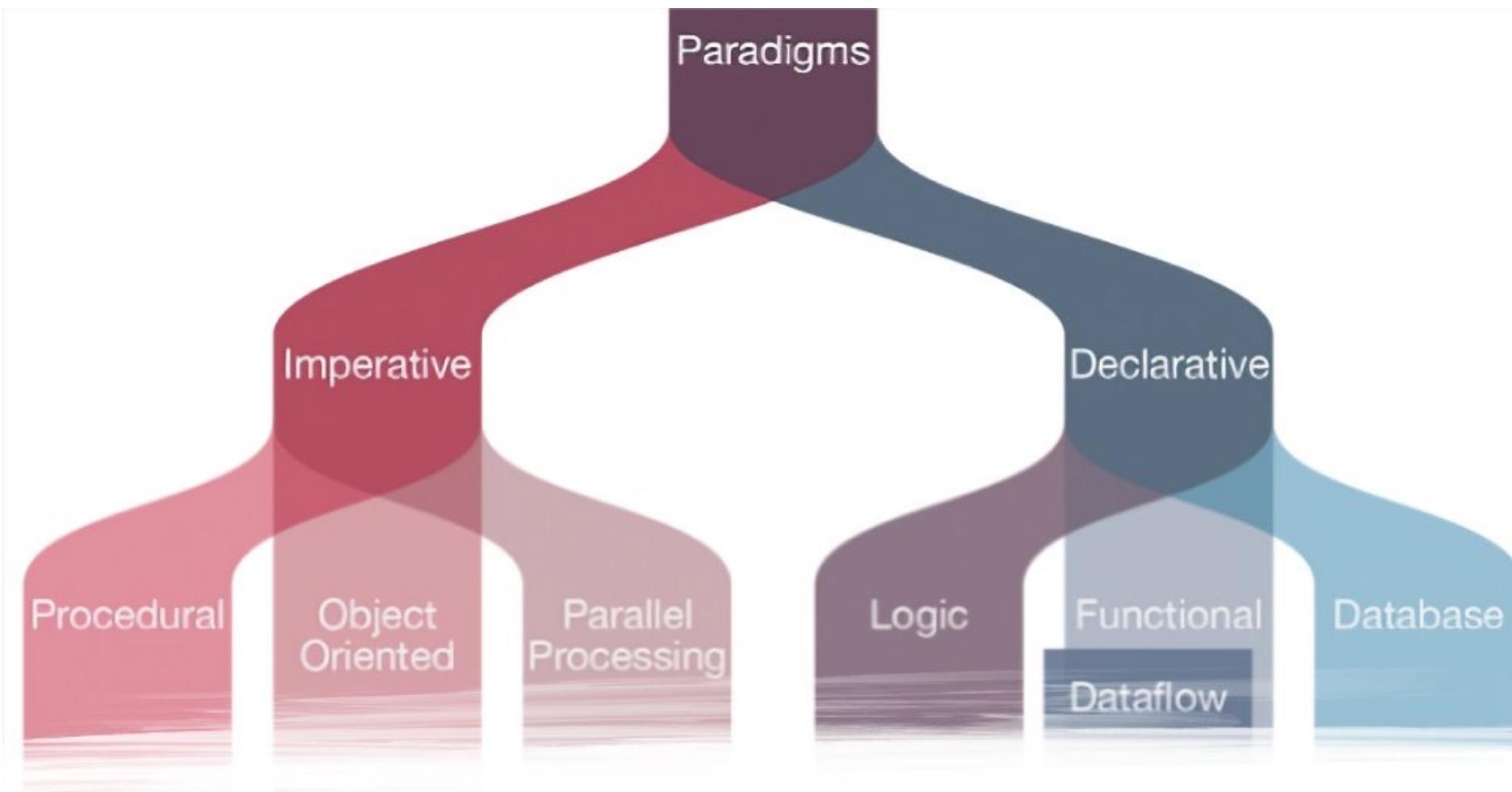


Scala



Northeastern University

Programming Paradigms



Source: <https://www.watelectronics.com/types-of-programming-languages-with-differences/>



Mutable vs. immutable data structures

Mutable

list

dictionary

Immutable

string

tuple

named tuple



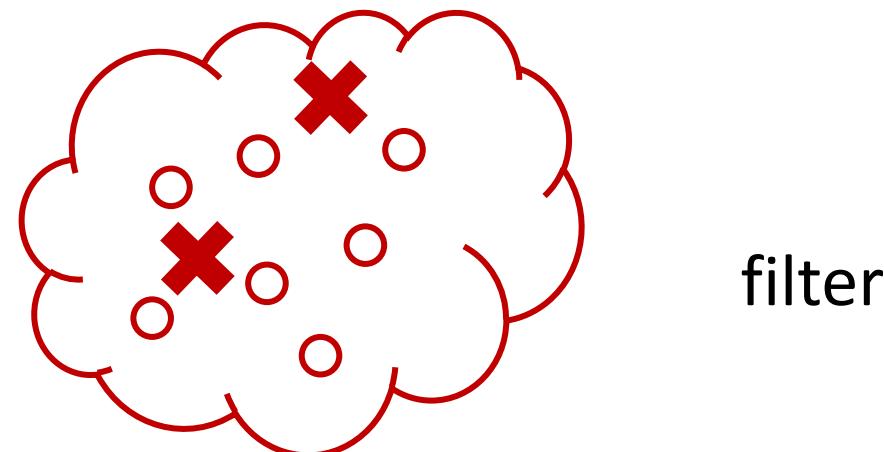
filter()

filter: Filter a collection according to which elements satisfy a function

filter(function, collection) → {elem ∈ Collection | function(elem) = True}

L = [1,2,3,4,5]

filter(lambda x: x%2 == 0, L) → [2,4] # or rather an iterator

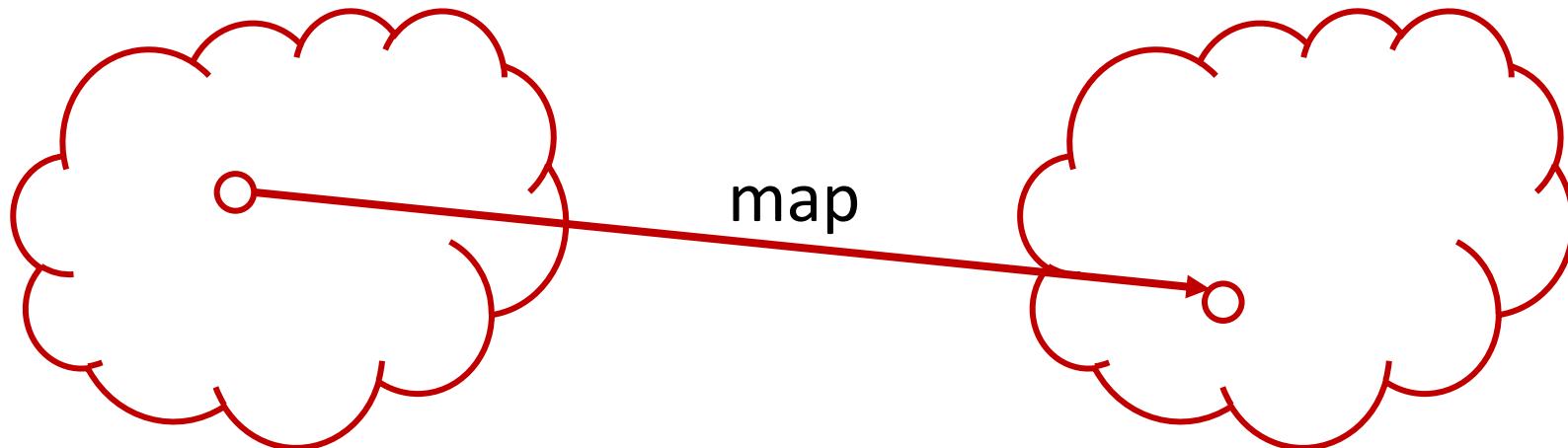


filter

map()

map: Maps elements of a collection to a new collection by applying a function to each element.

map(lambda x: x2, L) → [1,4,9,16,25]**



reduce()

reduce

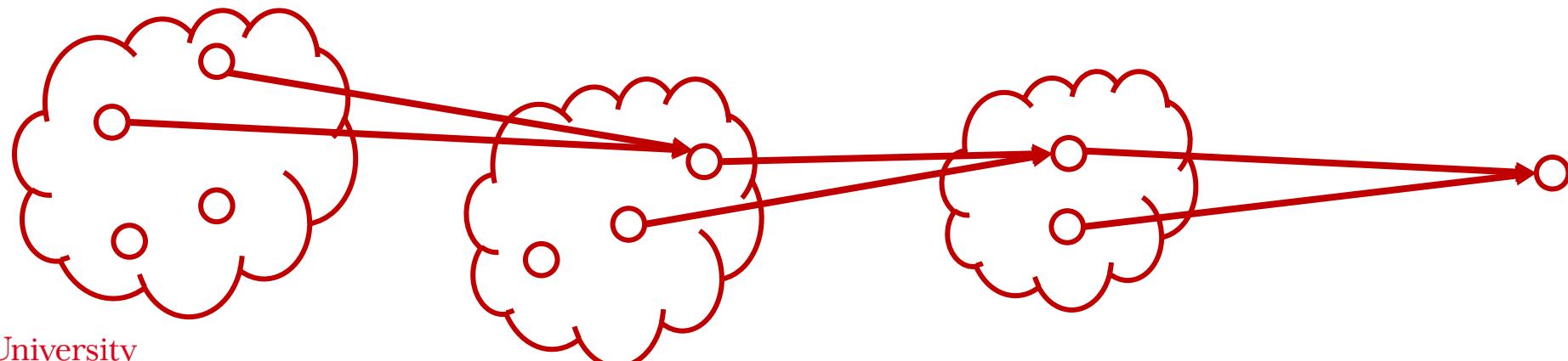
reduce: Apply a function of two arguments cumulatively to the items of a sequence, from left to right, so as to reduce the sequence to a single value..

L = [1,2,3,4,5]

reduce(function, collection, initial_value)

reduce(lambda x,y: x+y, L) → [3,3,4,5] → [6,4,5] → [10,5] → 15

NOTE: from functools import reduce



Intro to Parallel Programming



What is machine learning?

- Can we really make our machines (computers) learn?
- “Secret sauce” is **data, and lots of it**
- **Rather than programming expertise into our applications, we program them to learn from data**
- Build working machine-learning models then use them to make **remarkably accurate predictions**

Prediction

- Improve **weather forecasting** to save lives, minimize injuries and property damage
- Improve **cancer diagnoses and treatment regimens** to save lives
- Improve **business forecasts** to maximize profits and secure people’s jobs
- **Detect fraudulent credit-card purchases and insurance claims**
- Predict **customer “churn”**, what prices houses are likely to sell for, ticket sales of new movies, and anticipated revenue of new products and services
- Predict the **best strategies for coaches and players** to use to win more games and championships
- All of these kinds of predictions are happening today with machine learning.



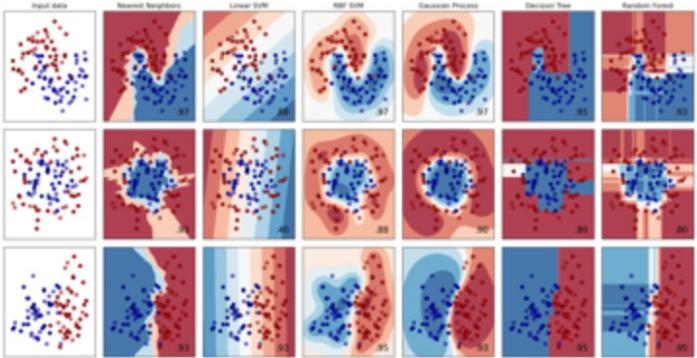
Machine Learning Approaches

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: [SVM](#), nearest neighbors, random forest, and [more...](#)

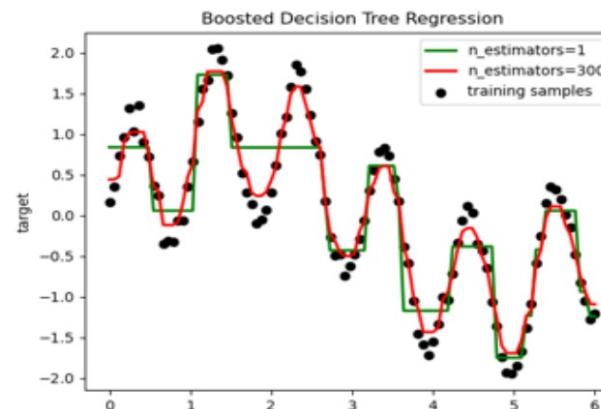


Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: [SVR](#), nearest neighbors, random forest, and [more...](#)

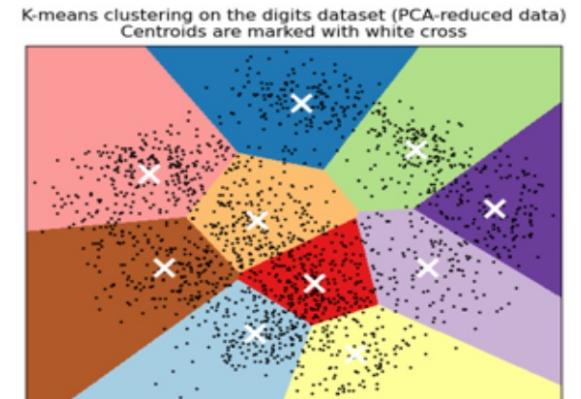


Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: [k-Means](#), spectral clustering, mean-shift, and [more...](#)



Popular Machine Learning Applications

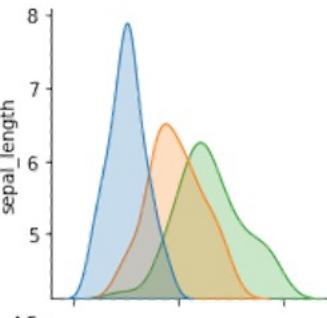
Anomaly detection	Data mining social media (like Facebook, Twitter, LinkedIn)	Predict mortgage loan defaults
Chatbots	Detecting objects in scenes	Natural language translation (English to Spanish, French to Japanese, etc.)
Classifying emails as spam or not spam	Detecting patterns in data	Recommender systems (“people who bought this product also bought...”)
Classifying news articles as sports, financial, politics, etc.	Diagnostic medicine	Self-Driving cars (more generally, autonomous vehicles)
Computer vision and image classification	Facial recognition	Sentiment analysis (like classifying movie reviews as positive, negative or neutral)
Credit-card fraud detection	Handwriting recognition	Spam filtering
Customer churn prediction	Insurance fraud detection	Time series predictions like stock-price forecasting and weather forecasting
Data compression	Intrusion detection in computer networks	Voice recognition
Data exploration	Marketing: Divide customers into clusters	



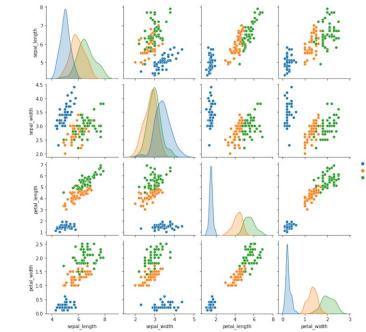
Machine Learning Recipe

```
5.1,3.8,1.6,0.2,Iris-setosa  
4.6,3.2,1.4,0.2,Iris-setosa  
5.3,3.7,1.5,0.2,Iris-setosa  
5.0,3.3,1.4,0.2,Iris-setosa  
7.0,3.2,4.7,1.4,Iris-versicolor  
6.4,3.2,4.5,1.5,Iris-versicolor  
6.9,3.1,4.9,1.5,Iris-versicolor  
5.5,2.3,4.0,1.3,Iris-versicolor
```

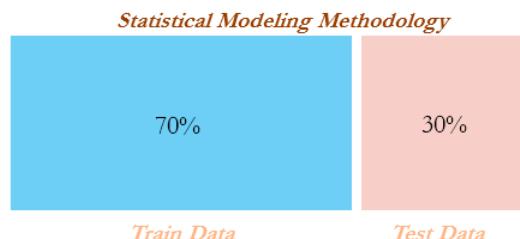
load
data



explore

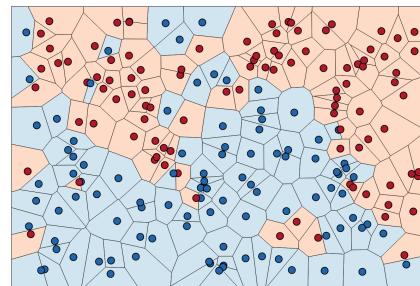


transform



training &
test split

Northeastern University



model
building

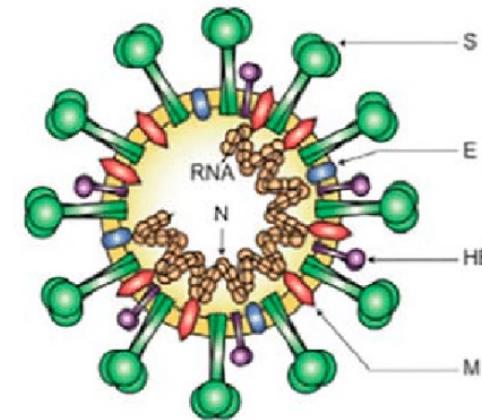
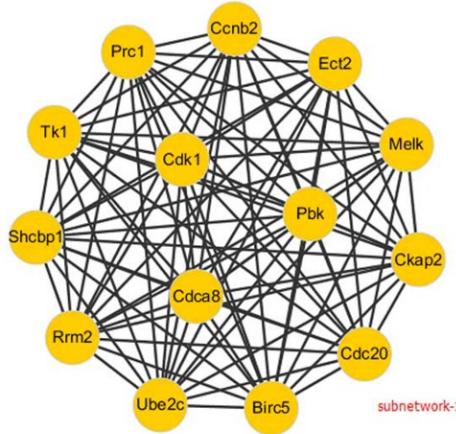
Some frequently used distance functions.	
Camerra :	$d(x, y) = \sum_{i=1}^m \frac{ x_i - y_i }{ x_i + y_i }$ (2)
Minkowsky :	$d(x, y) = \left(\sum_{i=1}^m x_i - y_i ^p \right)^{\frac{1}{p}}$ (3)
Chebychev :	$d(x, y) = \max_{i=1}^m x_i - y_i $ (4)
Euclidean :	$d(x, y) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$ (5)
Manhattan / city - block :	$d(x, y) = \sum_{i=1}^m x_i - y_i $ (6)

model
tuning

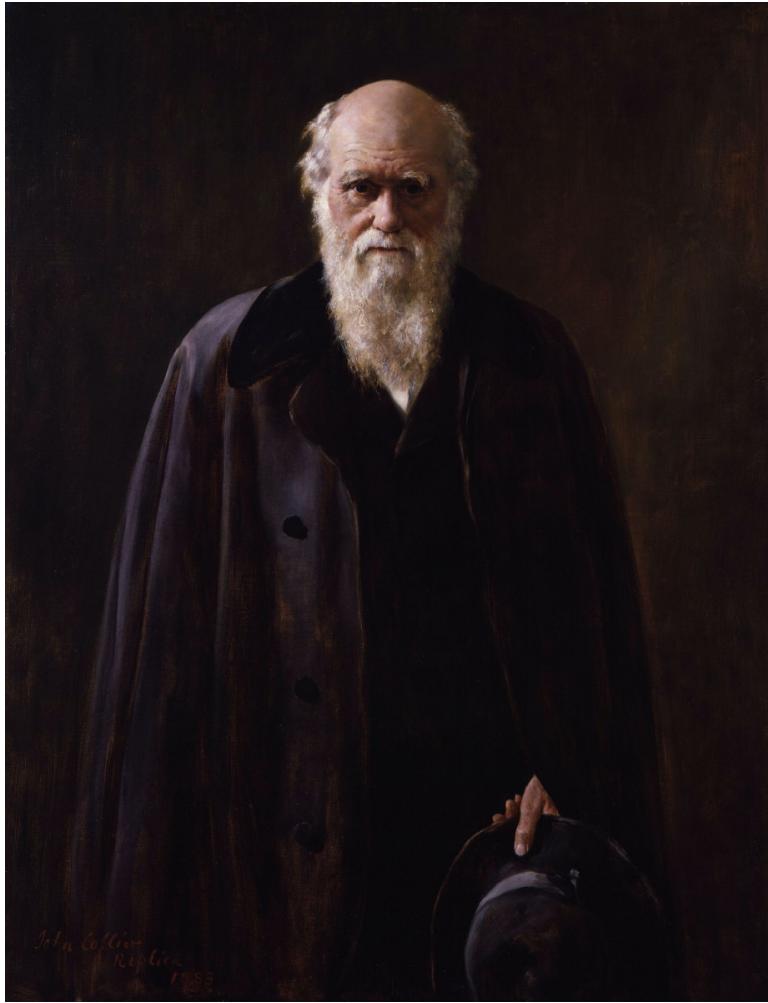


Data Science and Evolutionary Computing

Special Topics: Advanced Programming with Data



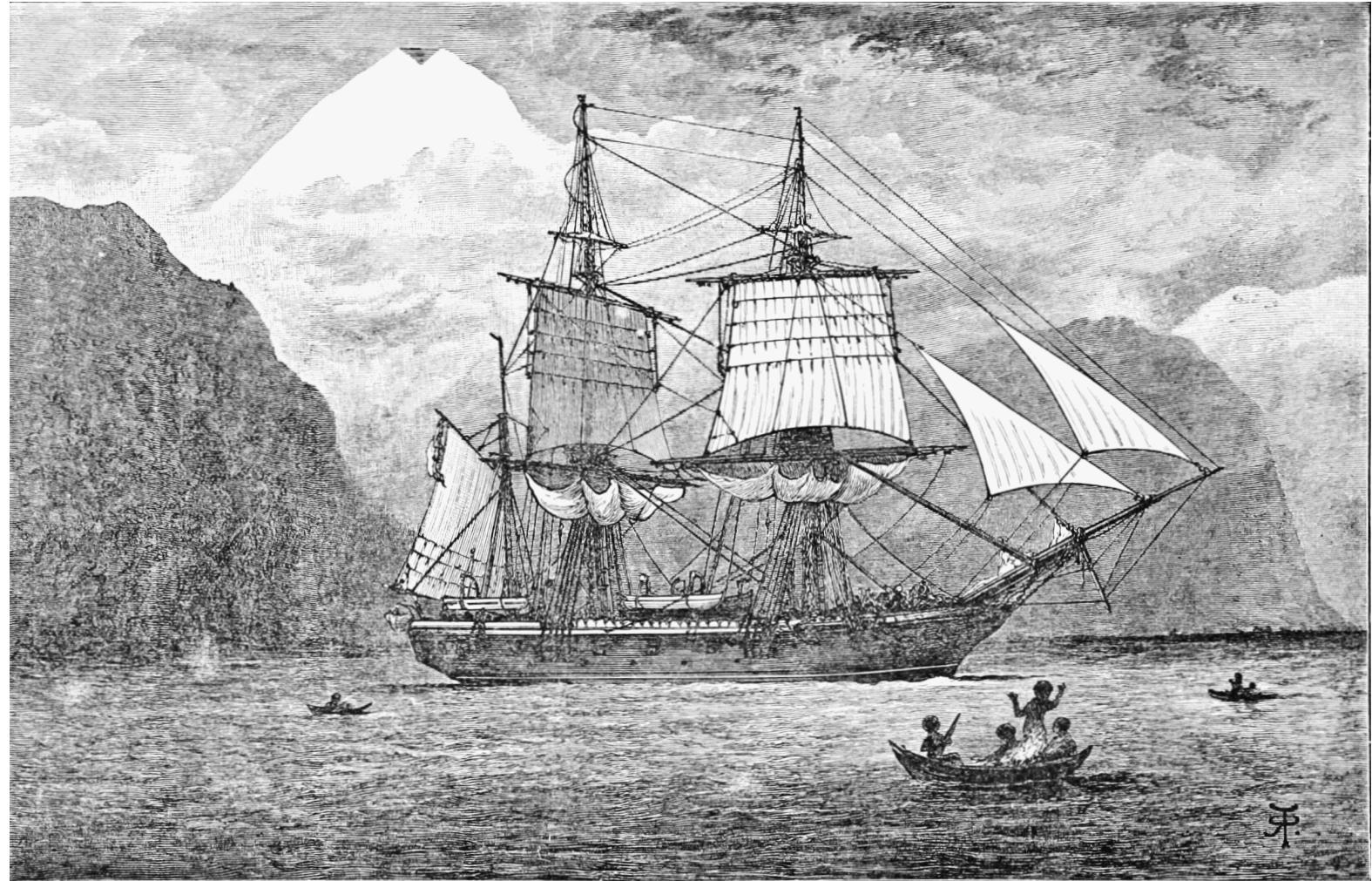
Charles Darwin



1809 - 1882

On the Origin of Species (1859)

Northeastern University



HMS Beagle

Voyage of the Beagle: 1831 – 1836

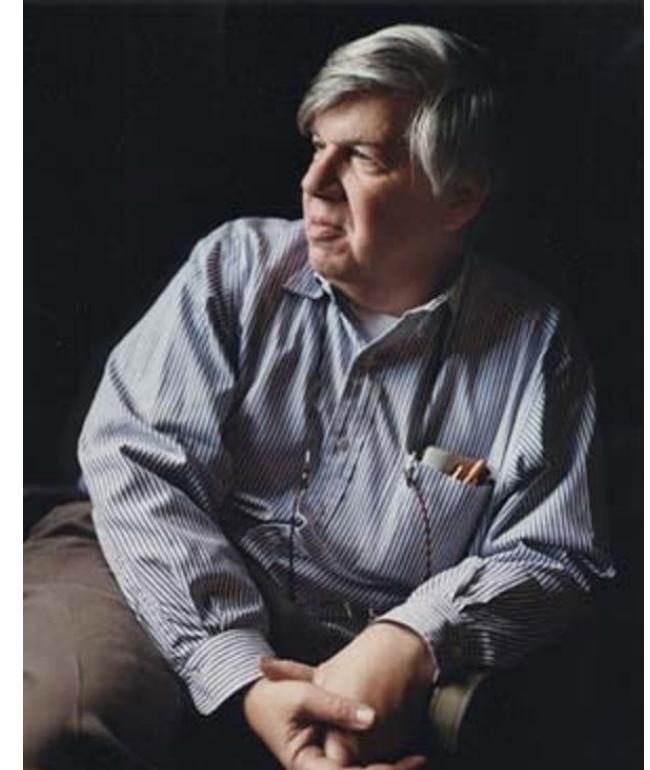


Natural Selection

[The] basis of natural selection is simplicity itself – two undeniable facts and an inescapable conclusion.

-- *Ever Since Darwin* (1977)

Stephen Jay Gould
Harvard Paleontologist
1941 - 2002



Variation and Inheritance

1. Organisms vary, and these variations are inherited (at least in part) by their offspring.



Population explosion

2. Organisms produce more offspring than can possibly survive.

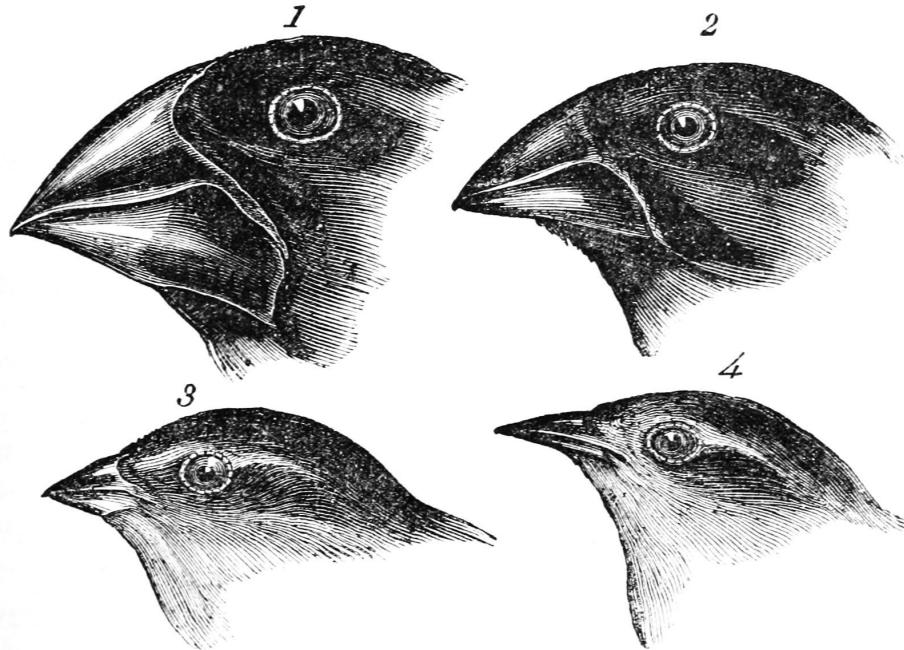


Survival of the Fittest

*3. On average,
offspring that vary
most strongly in
directions favored
by the environment
will survive and
propagate.*



Evolution by Natural Selection



1. *Geospiza magnirostris*.
3. *Geospiza parvula*.

2. *Geospiza fortis*.
4. *Certhidea olivacea*.

Survival of the
Fittest

Inheritance

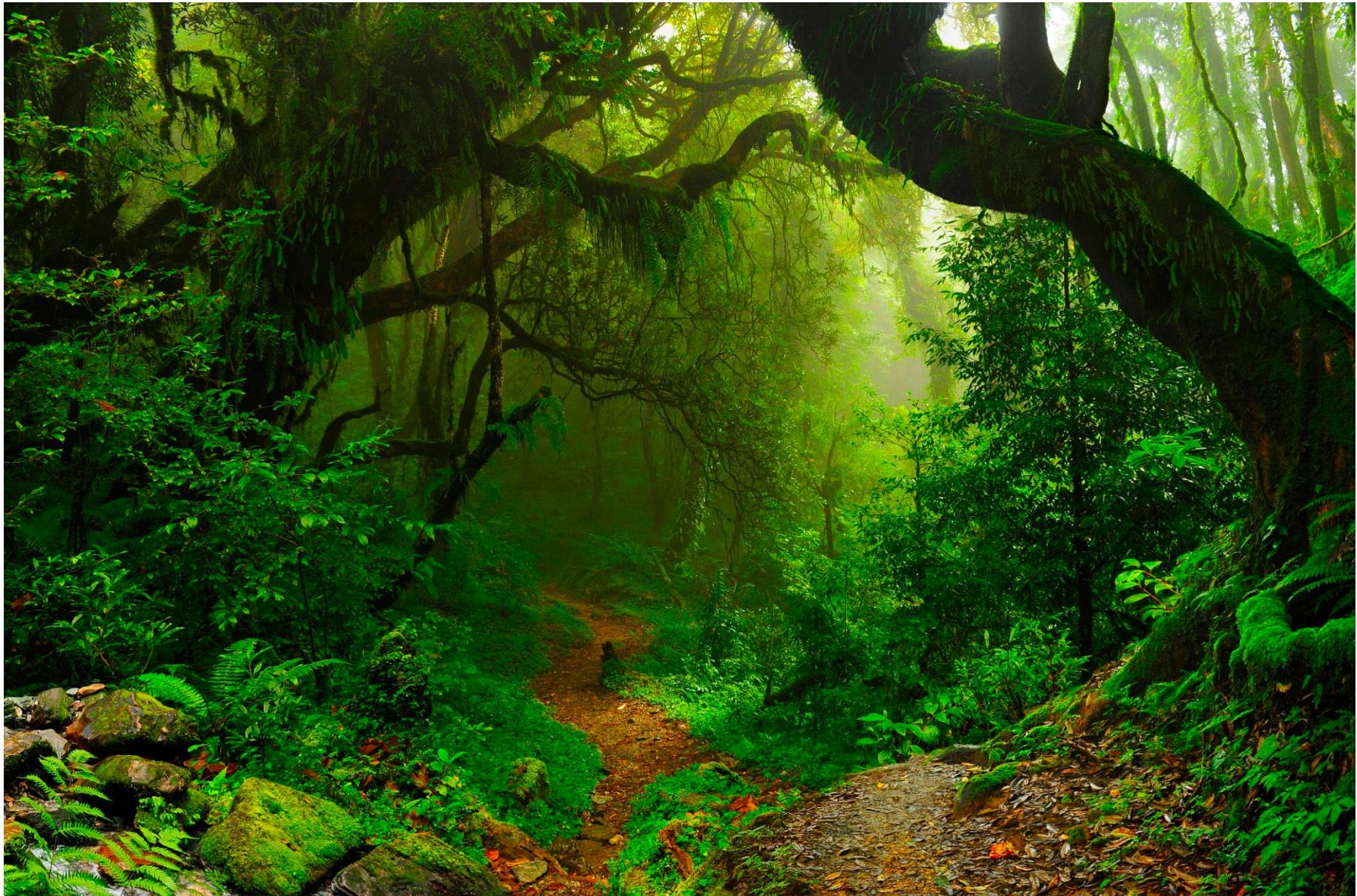
*Natural
Selection*

Overpopulation

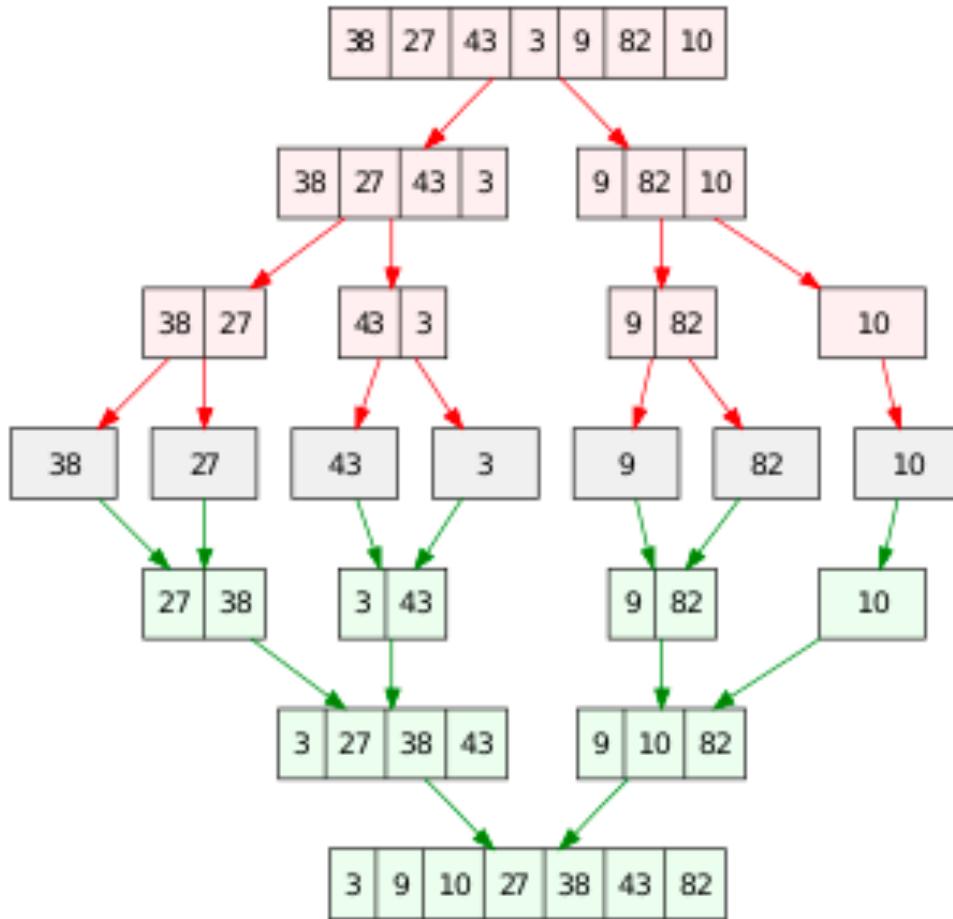


Biological systems are...

- Complex
- Diverse
- Adaptive
- Resilient
- Self-regulating (homeostatic)
- Synergistic (cooperative)
- Self-sustaining
- *CREATIVE!*



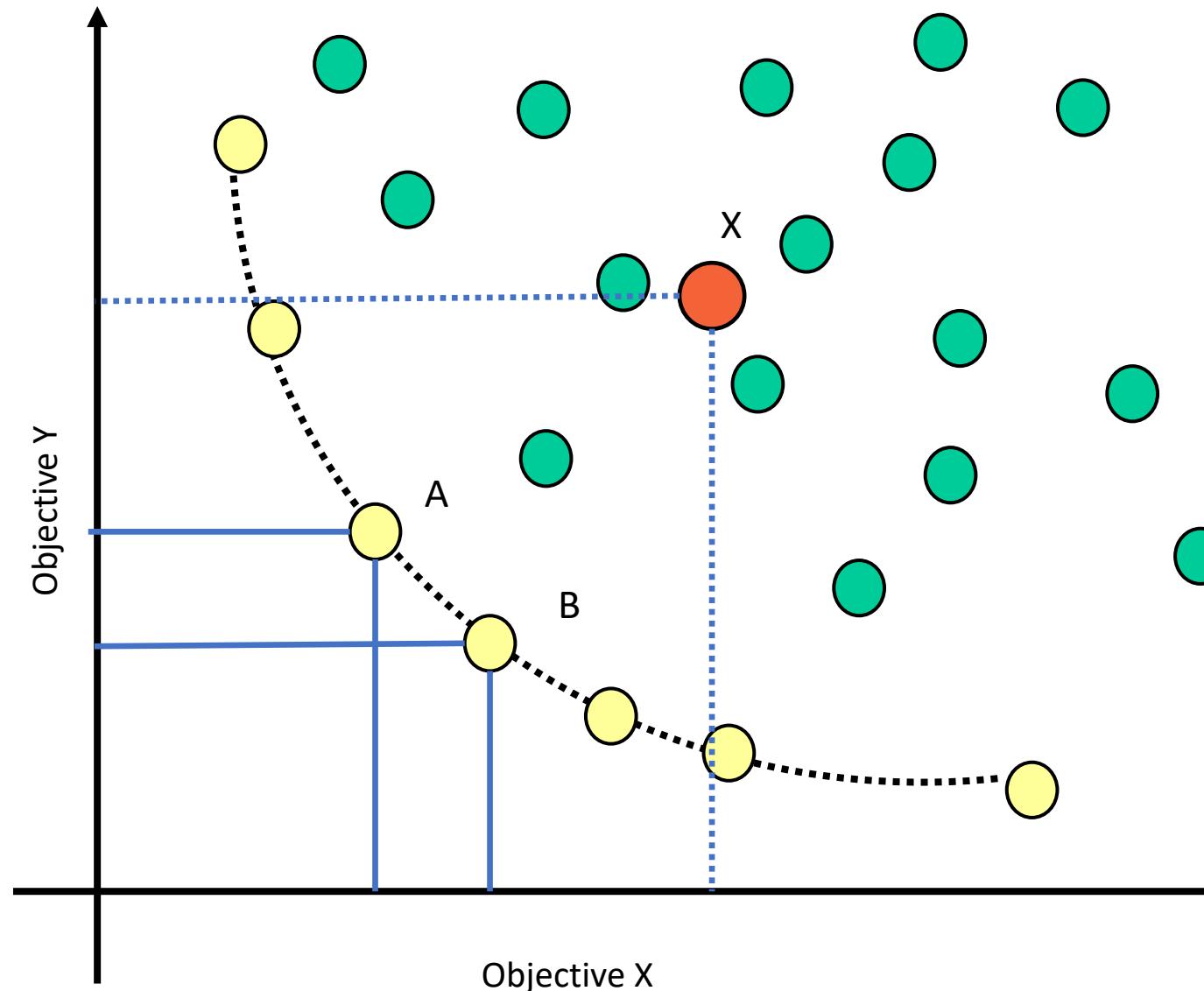
Most real-world problems are not like SORT



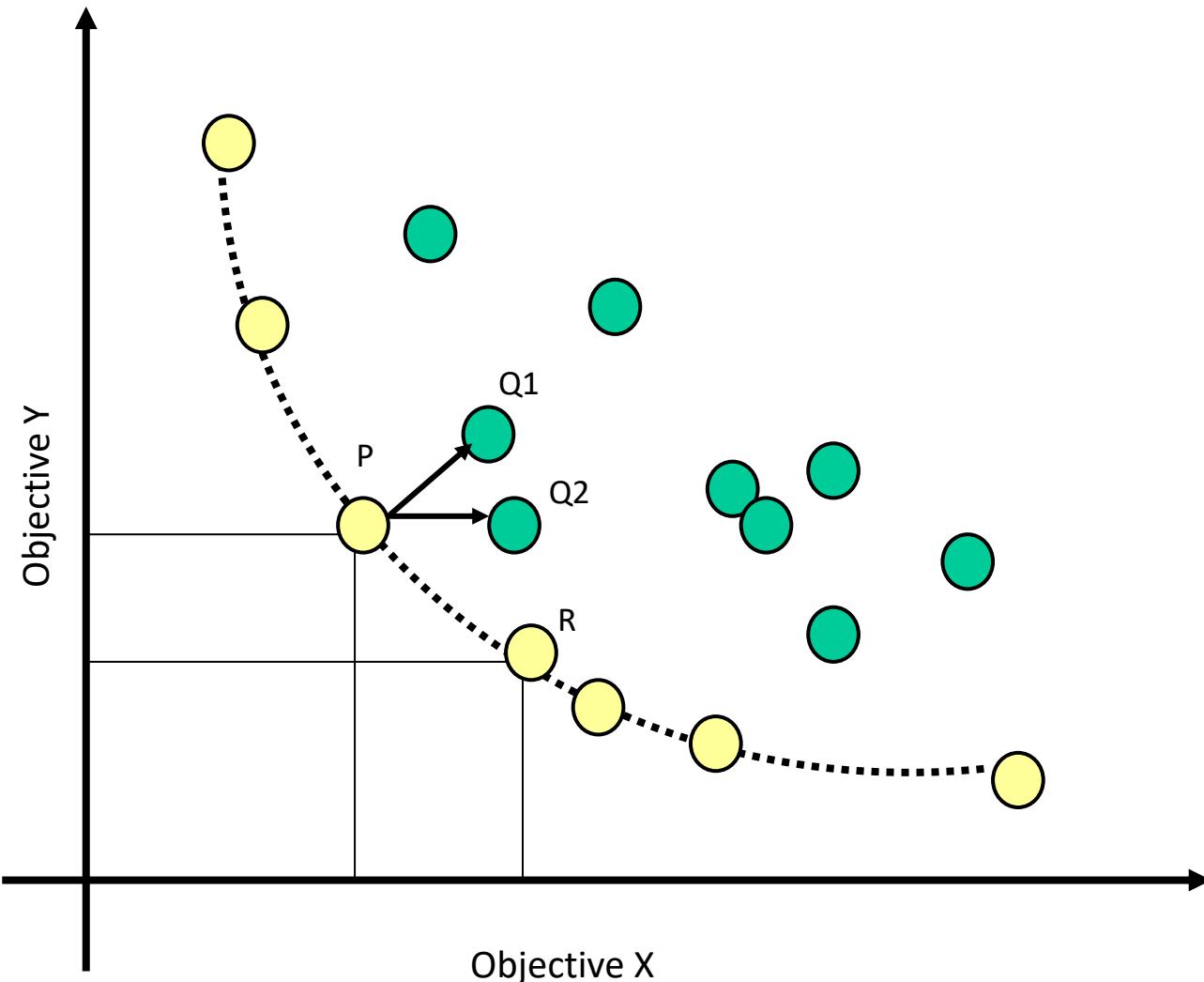
Usually, there is no single OPTIMAL solution...



Evolving Tradeoffs



...only tradeoffs.

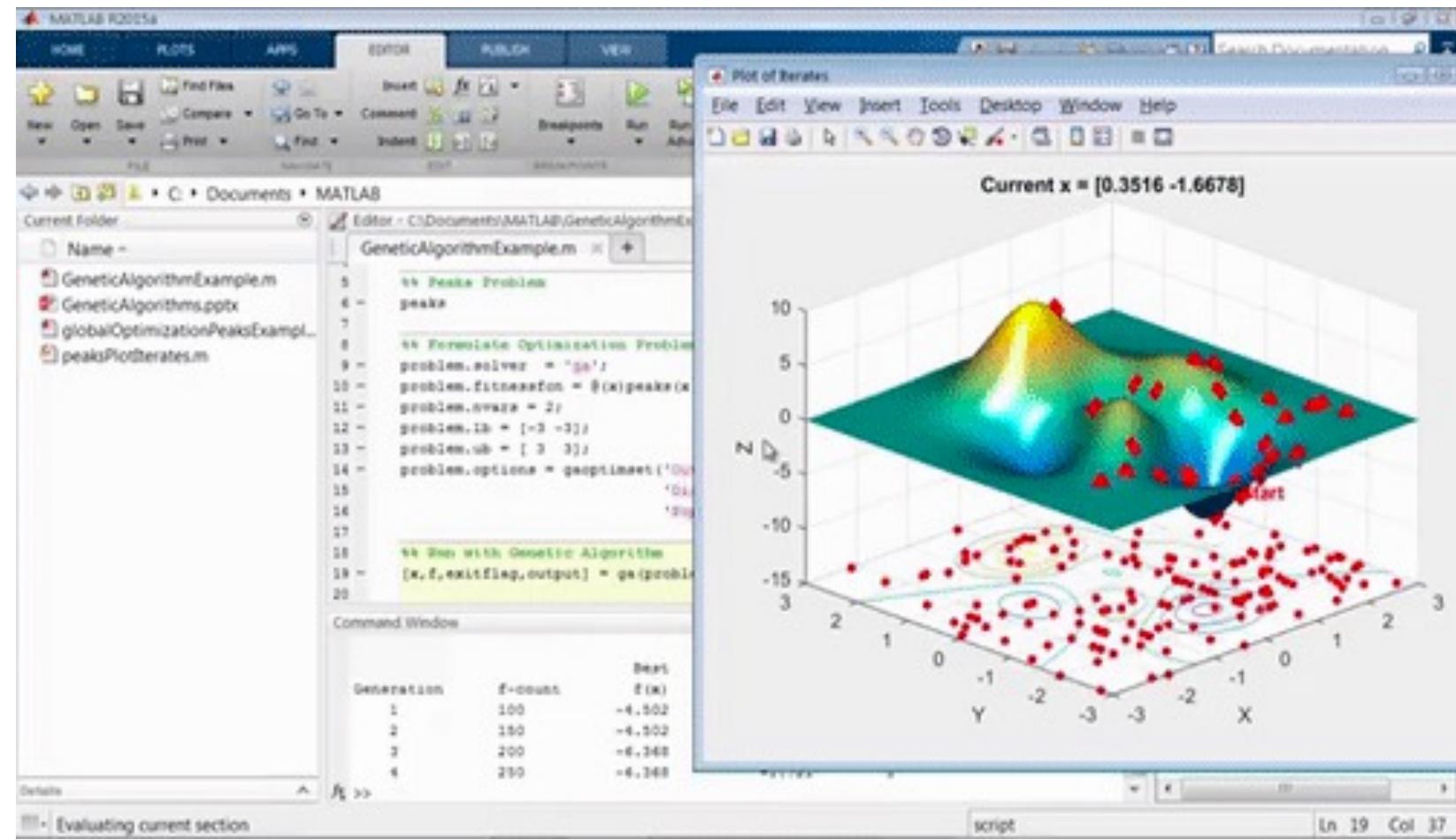
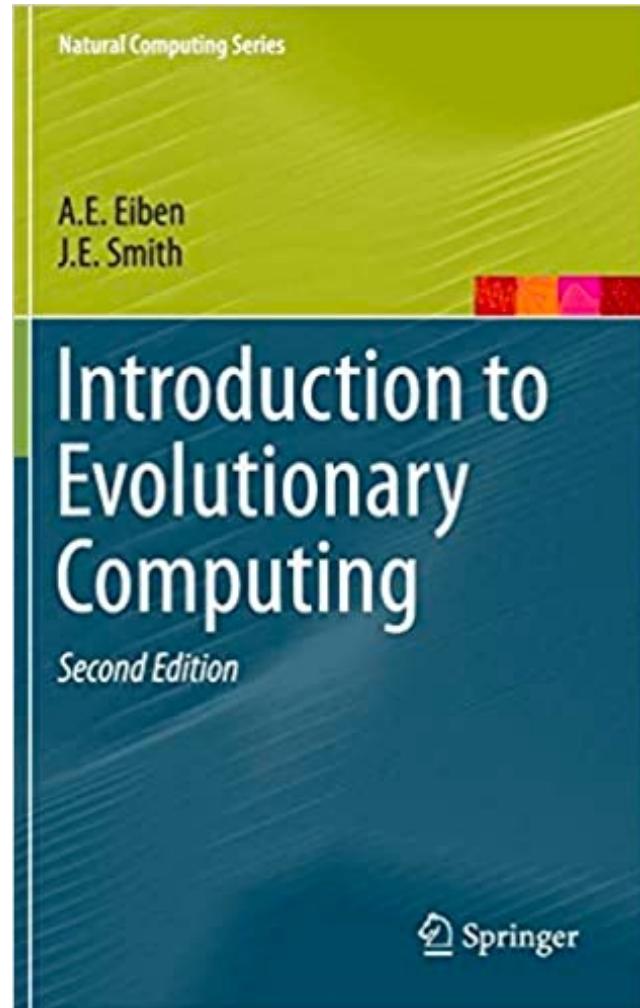


Effective management requires the careful consideration of competing alternatives.

- Peter Drucker



Algorithms that can evolve.



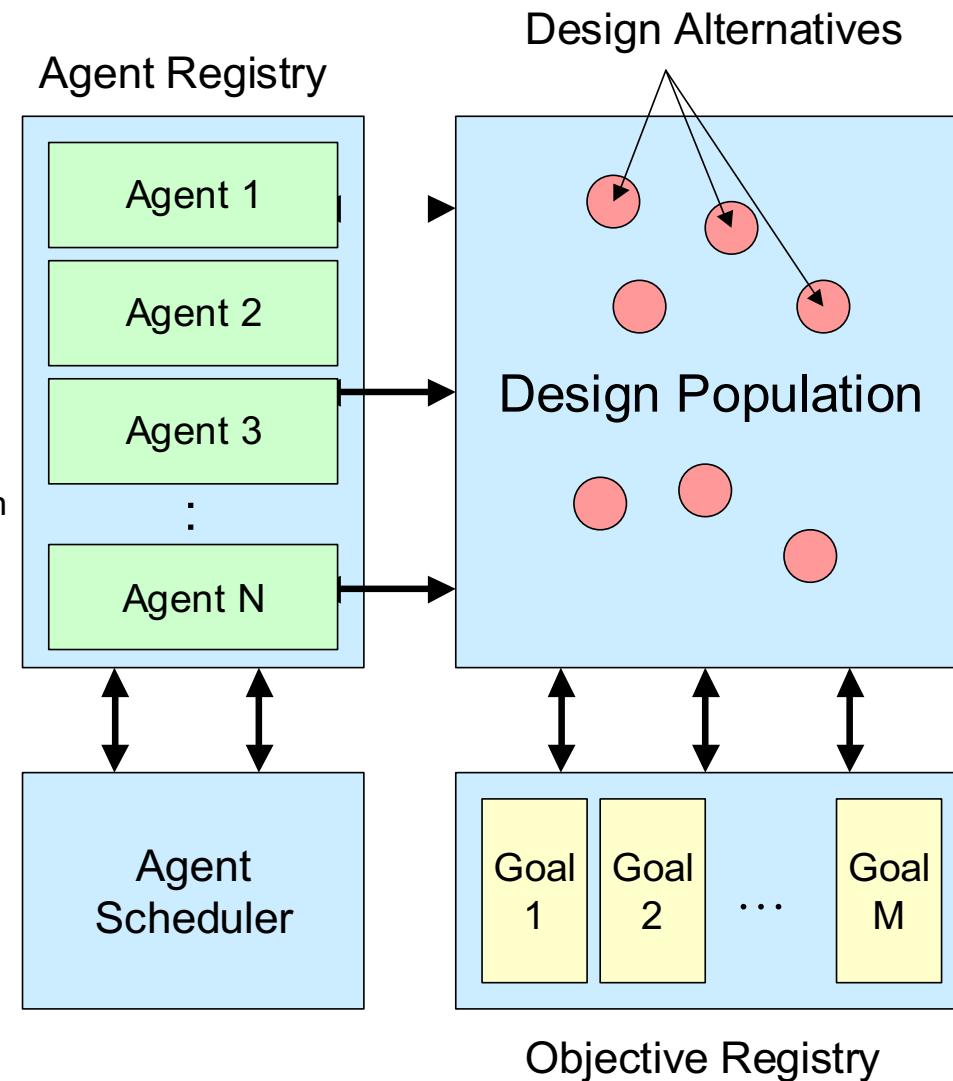
An architecture for multi-objective optimization using evolutionary computing

Classes of agents:

Creator Agent: Adds a design constructed from scratch

Improver Agent: Copies and then modifies an existing design

Destroyer Agent: Removes one or more unpromising designs



Sorting without **SORT**

1. Inherit Traits:

Copy a list

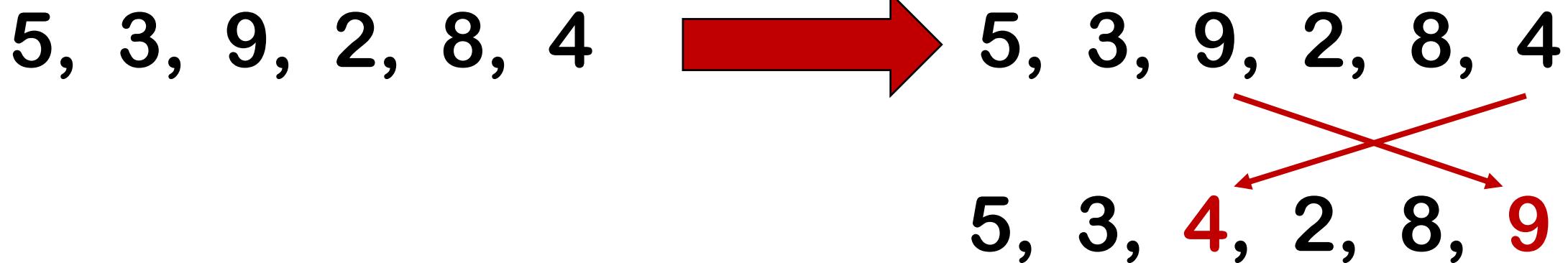
5, 3, 9, 2, 8, 4  5, 3, 9, 2, 8, 4



Sorting without **SORT**

2. Create variations:

Swap two elements (of the copy)



Sorting without **SORT**

3. Measure of fitness:

How *unsorted* are these lists? Count the **step downs**!

5, 3, 9, 2, 8, 4



$$2 + 7 + 4 = 13$$

5, 3, 4, 2, 8, 9



$$2 + 2 = 4$$



Sorting without **SORT**

3. Survival of the fittest:

How *unsorted* are these lists? Count the **step downs**!

5, 3, 9, 2, 8, 4



$$2 + 7 + 4 = 13$$



5, 3, 4, 2, 8, 9



$$2 + 2 = 4$$



k-Nearest Neighbor Learning

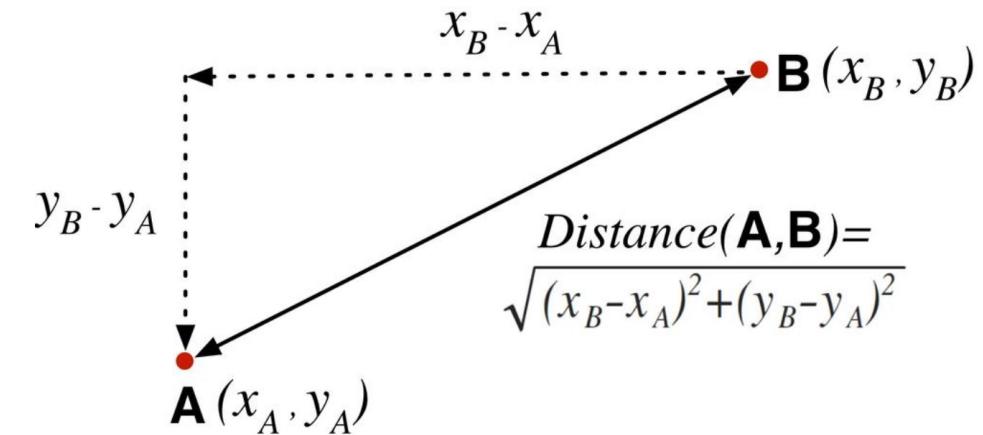
DS2000: Intermediate Programming with Data



Northeastern University

Euclidean Distance

Attribute	Person A	Person B
Age	23	40
Years at current address	2	10
Residential status (1=Owner, 2=Renter, 3=Other)	2	1



$$\begin{aligned}d(A, B) &= \sqrt{(23 - 40)^2 + (2 - 10)^2 + (2 - 1)^2} \\&\approx 18.8\end{aligned}$$

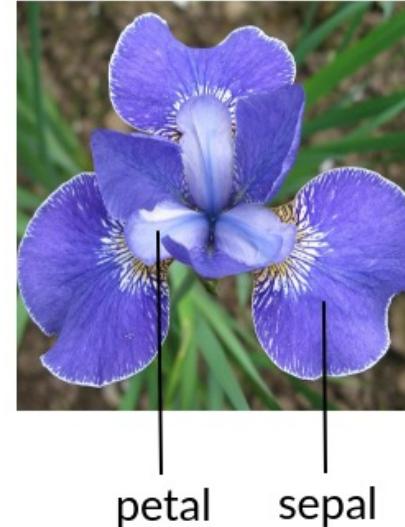


Euclidean Distance on Iris Dataset

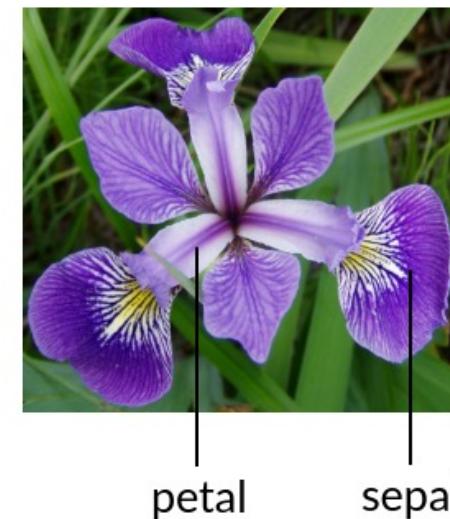
$$\text{diff} = \sqrt{(\Delta Slen)^2 + (\Delta Swid)^2 + (\Delta Plen)^2 + (\Delta Pwid)^2}$$

Slen = Sepal Length
Swid = Sepal Width
Plen = Petal Length
Pwid = Petal Width

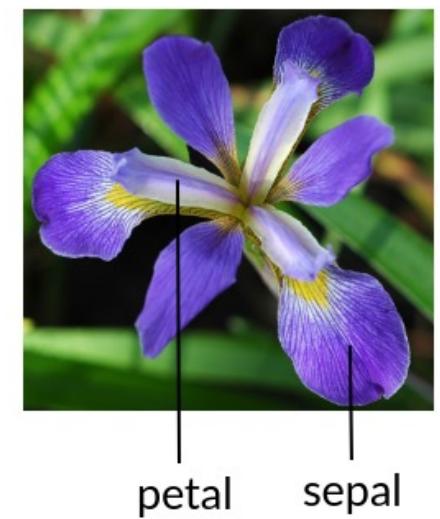
iris setosa



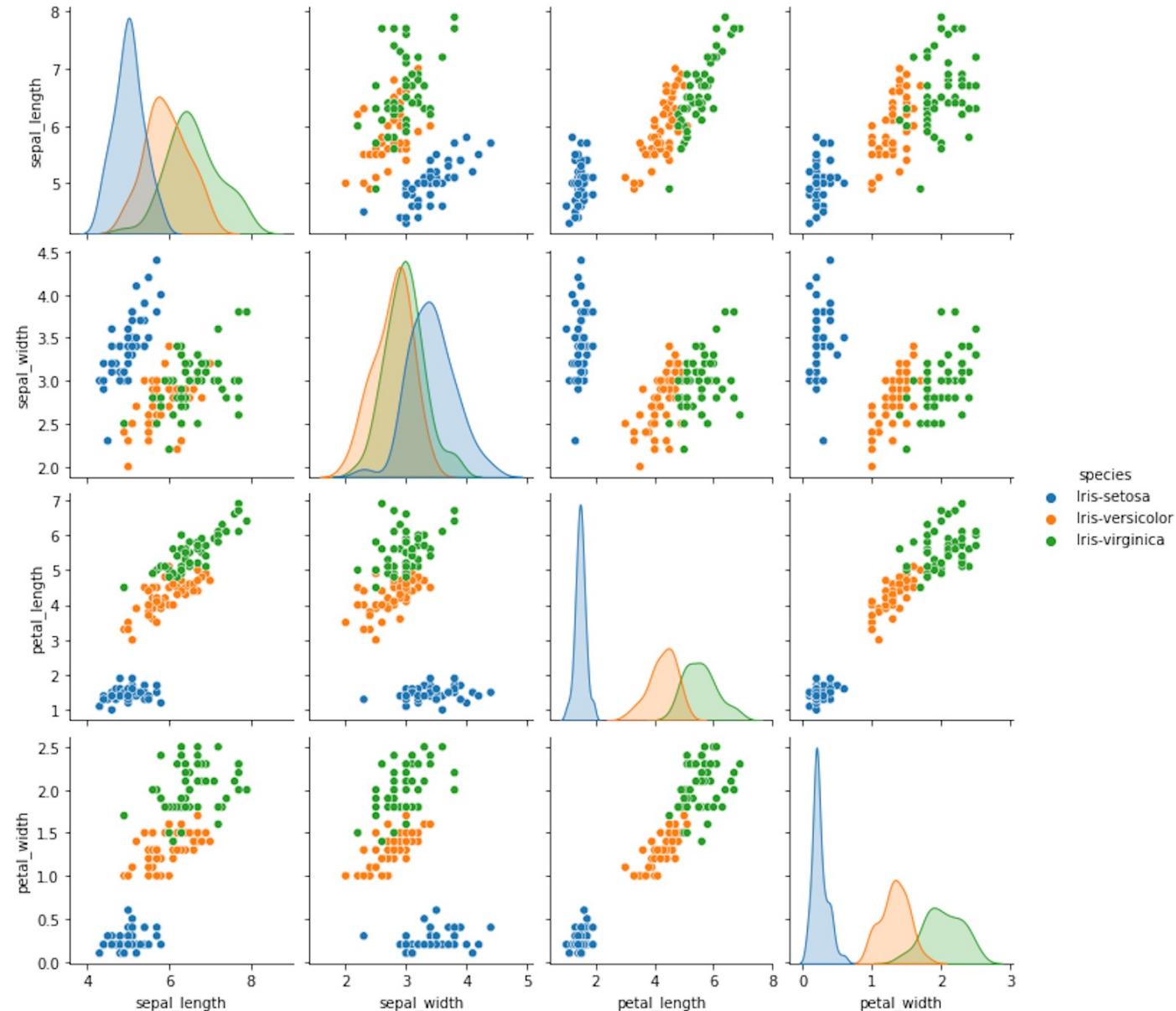
iris versicolor



iris virginica



Why 1-NN makes mistakes on the iris dataset



Iris Classification Report

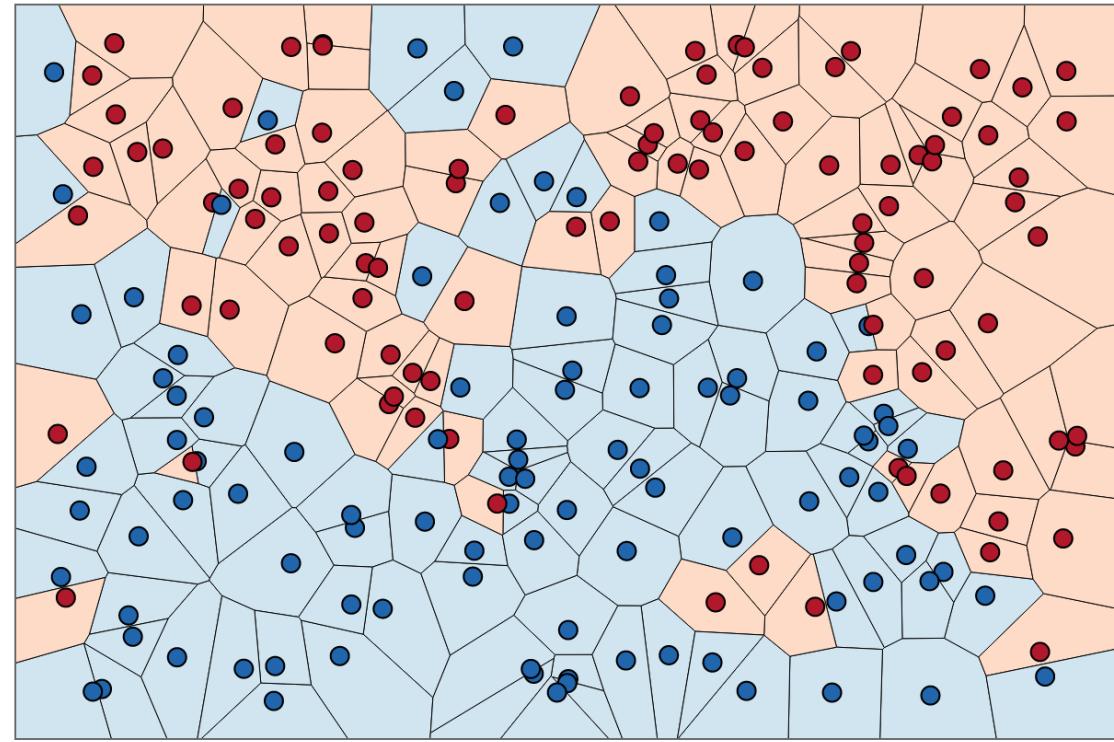
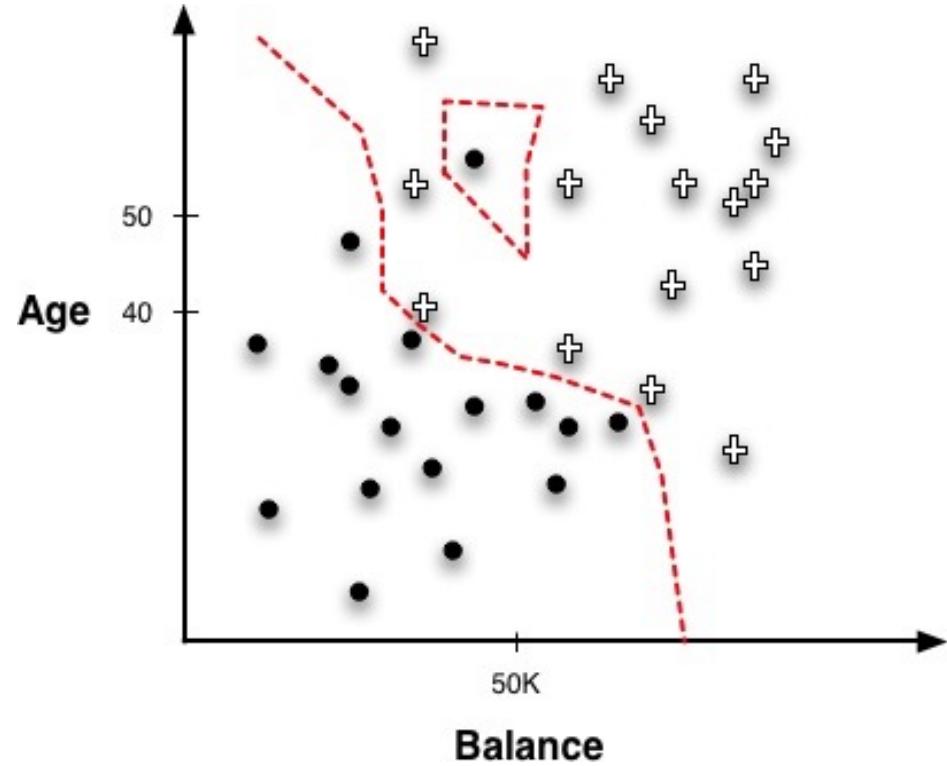
		actual		
		Setosa	Virginica	Versicolor
predicted	Setosa	50	0	0
	Virginica	0	47	3
	Versicolor	0	3	47

Classification Report:

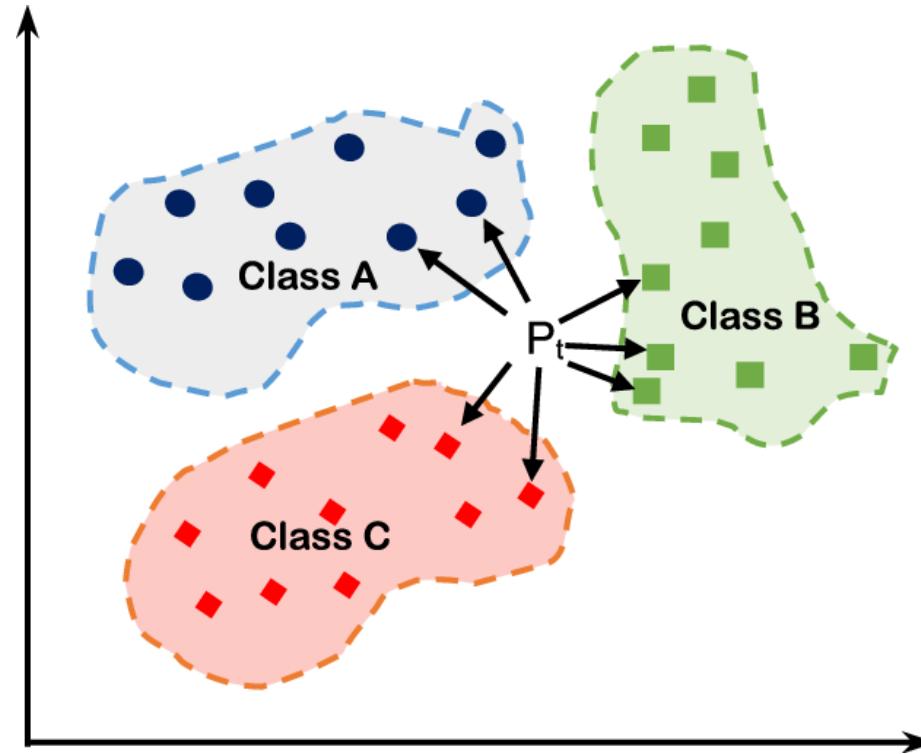
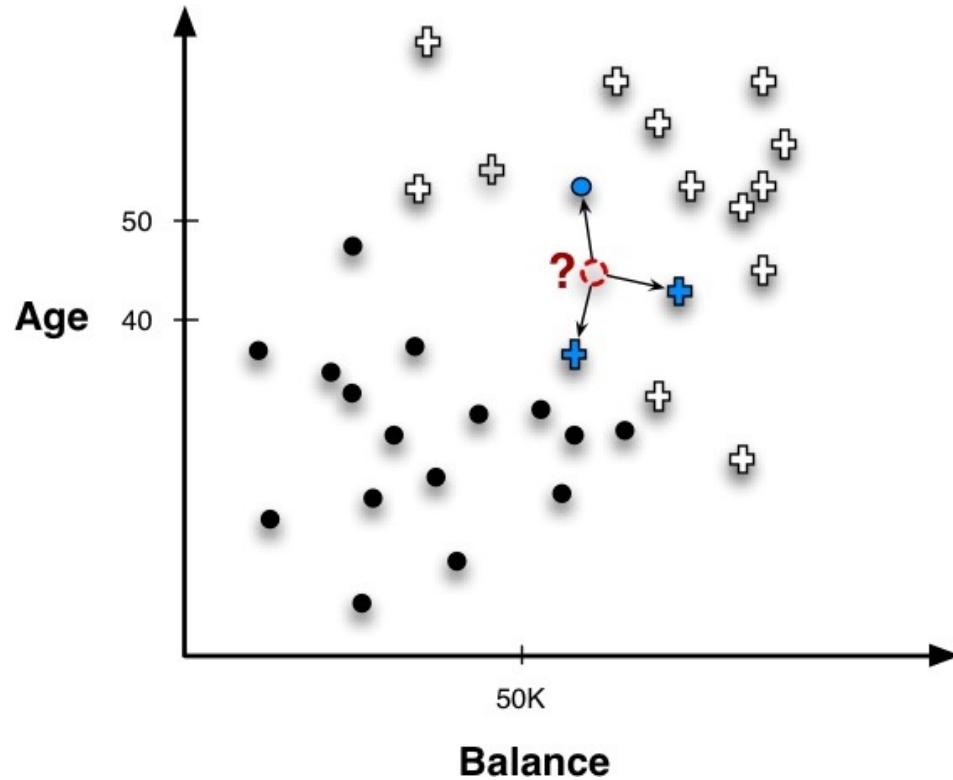
	precision	recall	f1-score	support
setosa	1.000	1.000	1.000	50
versicolor	0.940	0.940	0.940	50
virginica	0.940	0.940	0.940	50
accuracy			0.960	150
macro avg	0.960	0.960	0.960	150
weighted avg	0.960	0.960	0.960	150



Overfitting



Multiple classes



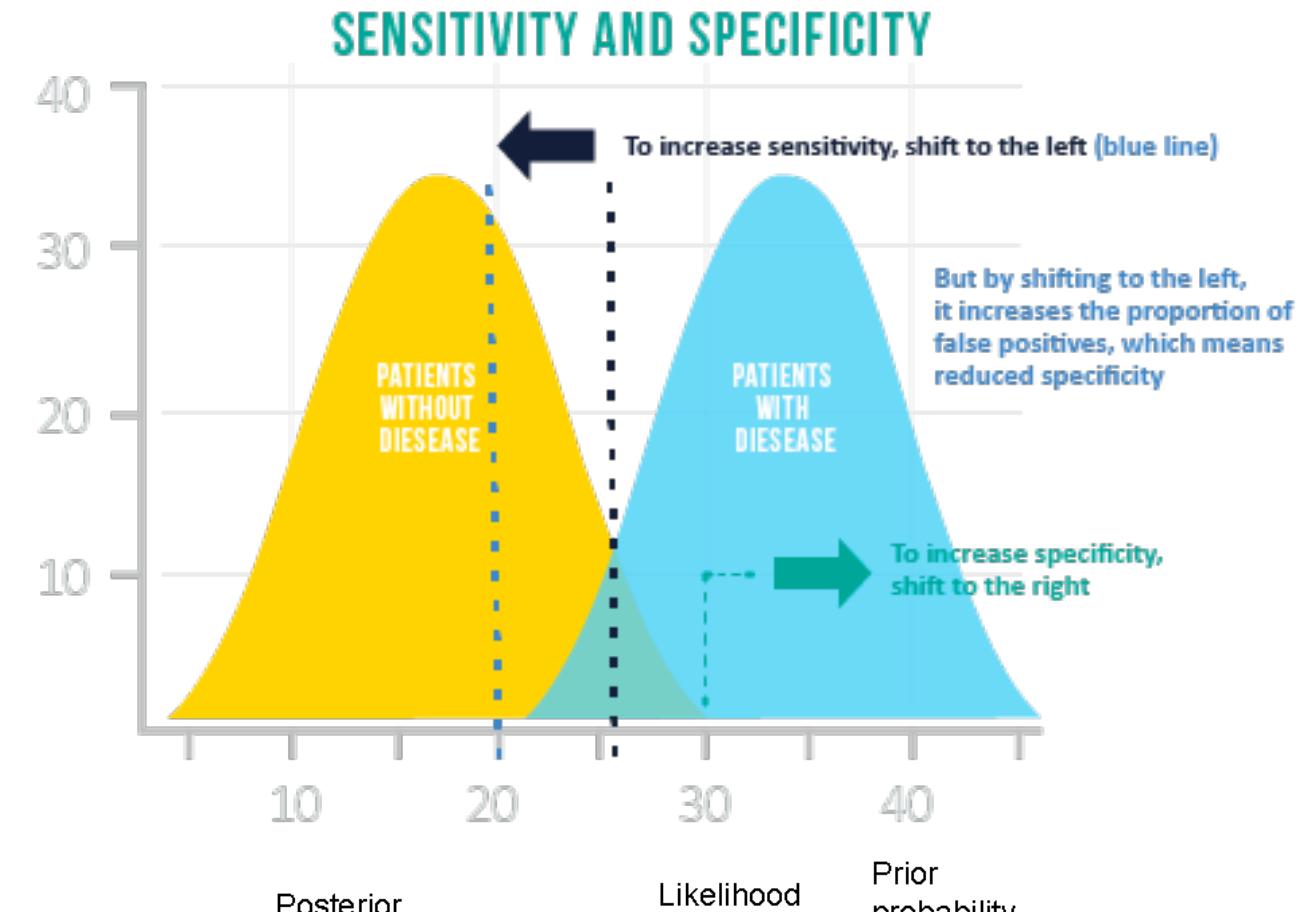
Source: <https://laptrinhx.com/k-nearest-neighbors-unlocked-454254569/>



Eventually we'll have a readily available test ...



But how will it impact public policy?



$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Measures of classification accuracy

10 photos, 7 dogs



Perfect Classification:

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	7	0
	Negative	0	3

Source: <https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2>



Imperfect Classification

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP) Type I Error
	Negative	2 (FN) Type II Error	2 (TN)

Recall: What proportion of ***actual positives*** are predicted positive?

$$\text{Recall} = TP / (TP + FN) = 5 / (5 + 2) = 71.4\%$$

Recall = Sensitivity



Imperfect Classification

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP) Type I Error
	Negative	2 (FN) Type II Error	2 (TN)

Precision: What proportion of ***predicted positives*** are truly positive?

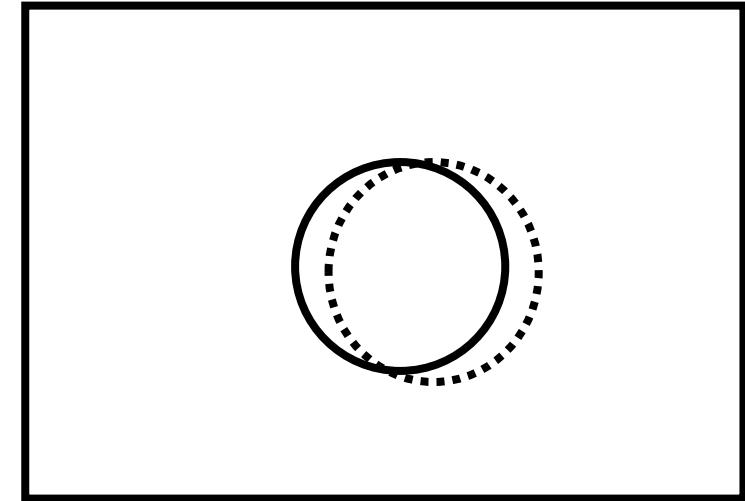
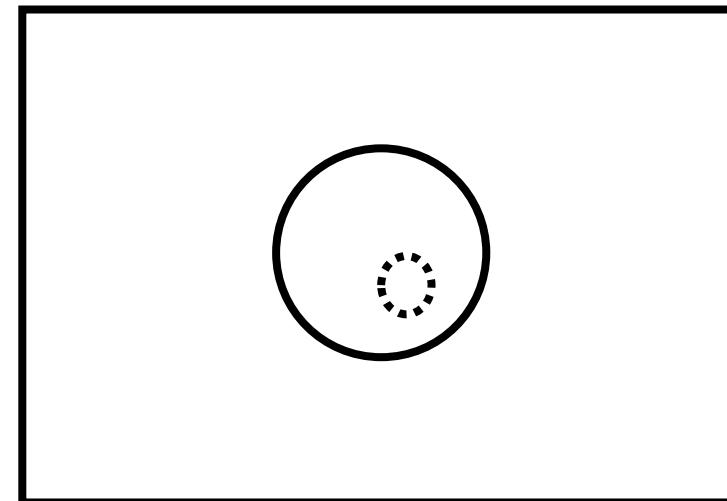
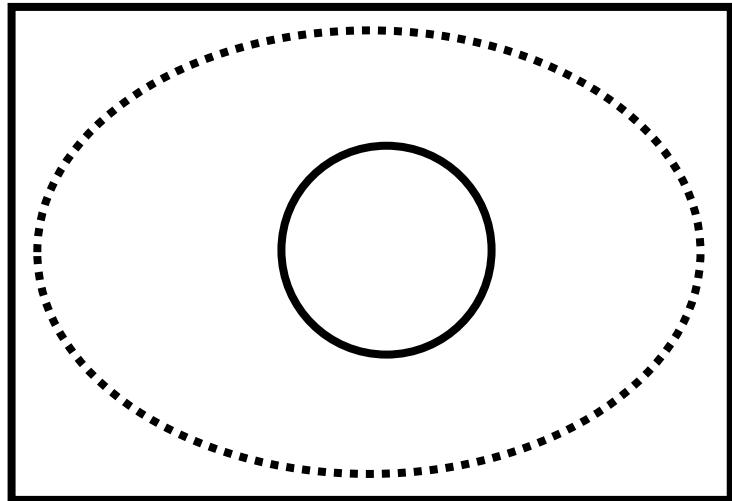
$$\text{Precision} = TP / (TP + FP) = 5 / (5 + 1) = 83.3\%$$

Specificity: What proportion of ***actual negatives*** are predicted negative?

$$\text{Specificity} = TN / (TN + FP) = 2 / (2 + 1) = 66.7\%$$



Recall/Precision and Sensitivity/Specificity



Perfect recall / sensitivity (TP/P):
Everyone truly sick is detected.

Poor precision (TP/(TP+FP)):
Few of the positives are really sick.

Poor specificity (TN/N):
Many healthy people are test positive.

Poor recall / sensitivity:
Many sick people aren't detected.

Perfect precision:
Every positive is truly sick.

Perfect specificity:
All healthy people test negative.

Good recall / sensitivity:
Most sick people are detected positive.

Good precision:
Most positives are *really* sick.

Good specificity:
Most healthy people test negative



General Distance Measure

$$\text{diff} = w_1 |\Delta A_1|^r + w_2 |\Delta A_2|^r + \dots + w_n |\Delta A_n|^r$$

where

ΔA_i is the difference with respect to feature i,

w_k is a feature weighting factor (0.0 to 1.0), and

r is an exponent (> 0.0)



Tweaking / modifying a solution

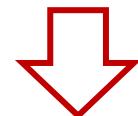
weights

r

k

1.0	1.0	1.0	1.0	2	1
-----	-----	-----	-----	---	---

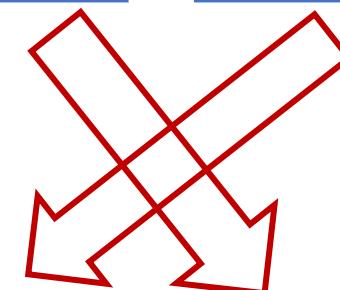
Mutate:



1.0	1.0	0.95	1.0	2	1
-----	-----	------	-----	---	---

0.20	0.90	1.00	0.45	2	5	0.38	0.29	1.00	0.93	1.5	3
------	------	------	------	---	---	------	------	------	------	-----	---

Crossover:



0.20	0.90	1.00	0.93	1.5	3	0.38	0.29	1.00	0.45	2	5
------	------	------	------	-----	---	------	------	------	------	---	---



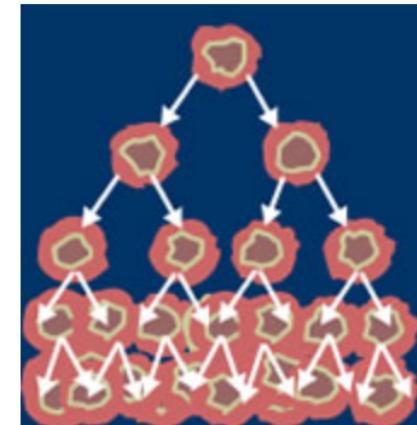
Predicting recurrence of breast cancer



Breast Cancer Wisconsin (Prognostic) Data Set

Download: [Data Folder](#), [Data Set Description](#)

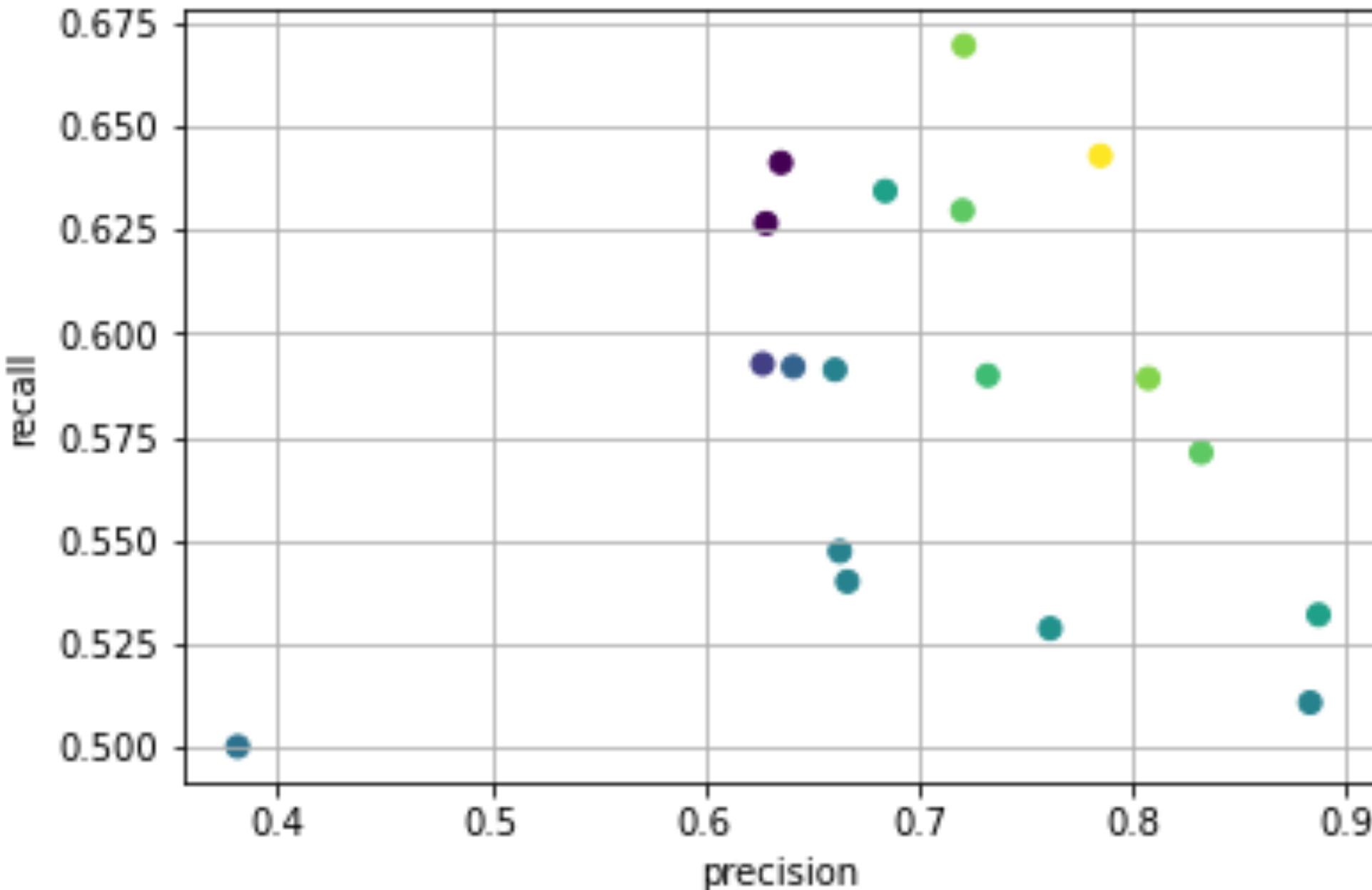
Abstract: Prognostic Wisconsin Breast Cancer Database



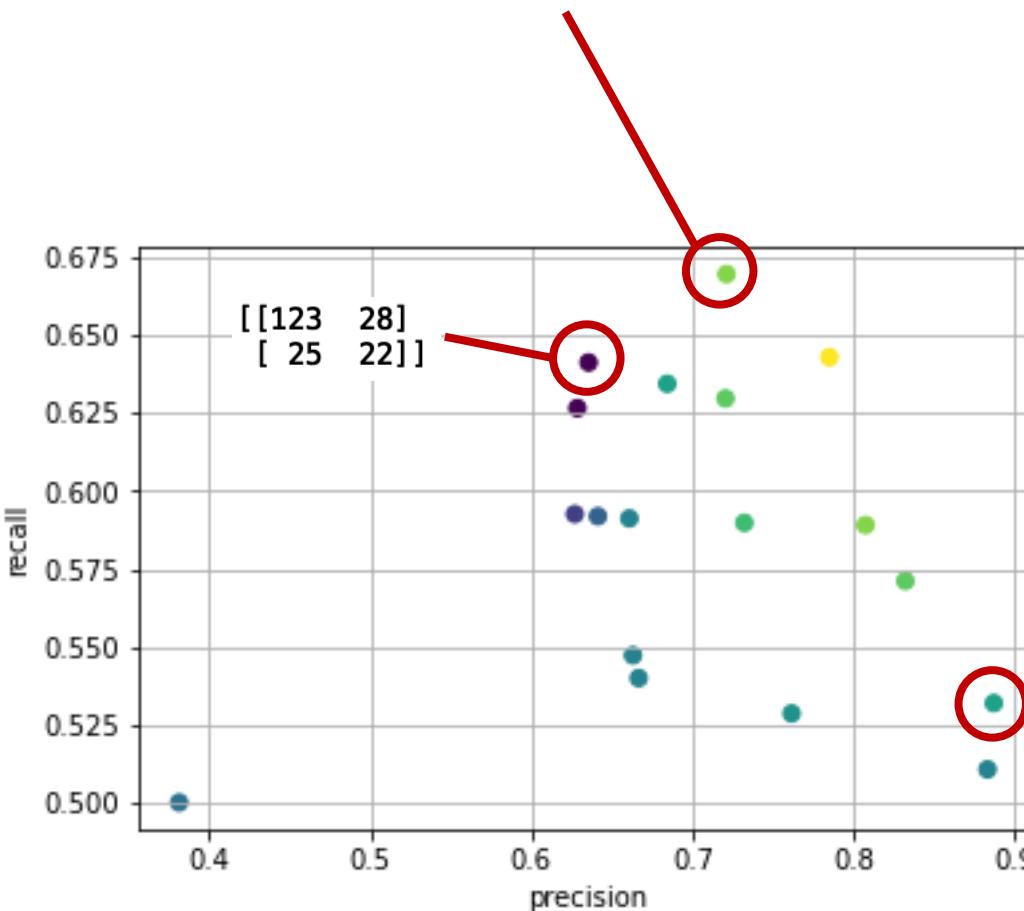
Data Set Characteristics:	Multivariate	Number of Instances:	198	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	34	Date Donated	1995-12-01
Associated Tasks:	Classification, Regression	Missing Values?	Yes	Number of Web Hits:	224238







actual	predicted		R=1.7, K=3, Features = 3			
	N	R	precision	recall	f1-score	support
N	[138 13]		0.83636	0.91391	0.87342	151
R	[27 20]		0.60606	0.42553	0.50000	47
accuracy				0.79798		198
macro avg	0.72121	0.66972		0.68671		198
weighted avg	0.78170	0.79798		0.78478		198



Recall: What fraction of the recurring cancers are we detecting (predicted to recur)?
 $(20 / 47 = 0.42553)$

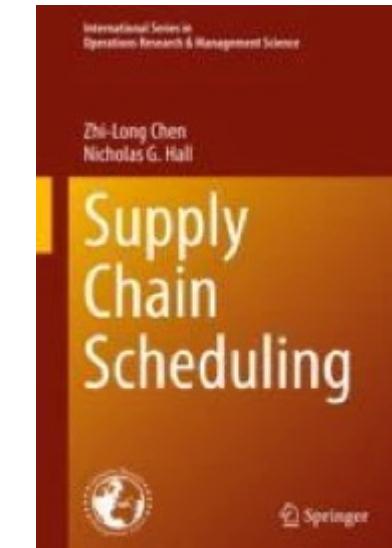
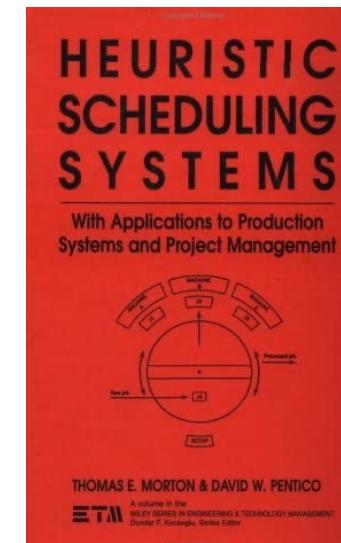
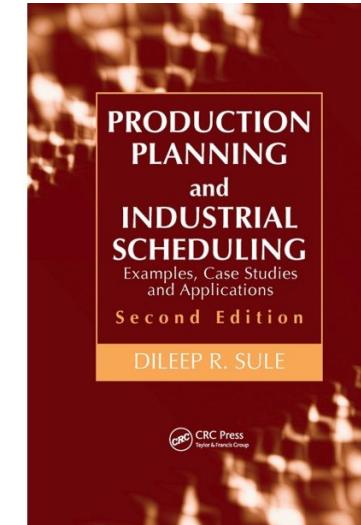
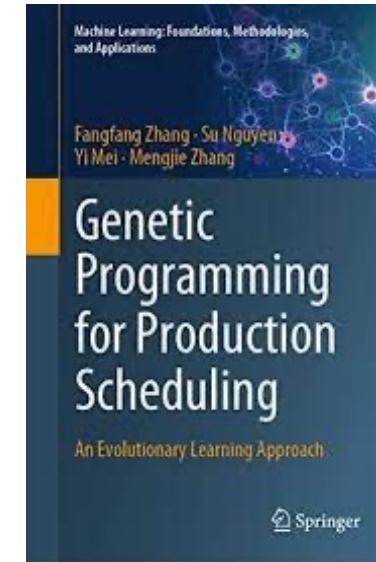
Precision: What fraction of those cancers predicted to recur actually recurred?
 $(20 / 33 = 0.60606)$

actual	predicted		R=2.3, K=17, Features = 1			
	N	R	precision	recall	f1-score	support
N	[151 0]		0.77436	1.00000	0.87283	151
R	[44 3]		1.00000	0.06383	0.12000	47
accuracy					0.77778	198
macro avg	0.88718	0.53191		0.49642		198
weighted avg	0.82792	0.77778		0.69413		198

R=2.3, K=17, Features = 1

		predicted					
		N	R				
actual	N	[[151 0]					
	R	[44 3]]					
				precision	recall	f1-score	support
		N	0.77436	1.00000	0.87283		151
		R	1.00000	0.06383	0.12000		47
	accuracy					0.77778	198
	macro avg		0.88718	0.53191	0.49642		198
	weighted avg		0.82792	0.77778	0.69413		198

Machine Scheduling



Northeastern University

Education: Faculty/Course, TA/Lab

		Section #:	1	2	3	5	4	6	8	9	10	11	13	14	15	16	17	18	20
	Instructor:	Rachlin	Sathyaranayara	Mosca	Zhang	Zhang	Luo	Gillani	Luo	Uslu	Gillani	Strange	Mosca	Strange	Mosca	Luo	Qu	Luo	
	Day/Time:	R 1145-125	W 950-1130	W 950-1130	W 1145-125	W 950-1130	W 1145-125	W 250-430	W 250-430	W 250-430	W 440-630	R 950-1130	R 950-1130	R 1145-125	R 1145-125	R 1145-125	R 250-430	R 250-430	
	Location:	ONLINE	WVH 212	WVH 210B	WVH 210A	KA 005	WVH 212	WVH 210A	WVH212	WVH 210B	WVH 210A	WVH 210B	WVH 212	WVH 210A	WVH 212	WVH 210B	WVH 212	WVH 210B	
	# Students:	36	35	19	35	19	40	19	19	36	35	19	36	19	19	19	39	35	479
TA	MAX SECTIONS	DS	Health	DS	SocSci	SocSci	Business	Business	Business	SocSci	Business	DS	DS	DS	DS	Business	SocSci	Business	
8 Browne L	2	U	W	P	U	ASSIGN	U	U	U	W	U	U	U	U	U	U	ASSIGN	W	2
9 Cassway R	1	ASSIGN	U	U	U	U	U	U	W	W	U	U	U	P	W	W	U	U	1
10 Chandra D	1	U	ASSIGN	P	U	P	U	U	U	U	U	U	P	P	U	U	U	U	1
11 Chandrashekhar S	2	W	W	U	U	U	U	U	U	U	U	U	P	U	U	U	U	U	2
12 Chenthikannan A	2	W	ASSIGN	W	U	U	U	U	U	U	U	U	U	U	U	U	P	U	2
13 Duzgun Z	1	ASSIGN	U	U	U	U	U	U	U	U	U	U	U	W	W	W	U	U	1
14 Gong J	2	U	U	U	U	U	U	U	U	U	U	U	U	W	W	W	P	ASSIGN	2
15 He C	1	W	U	U	U	U	U	U	U	U	U	U	U	P	ASSIGN	W	U	U	1
16 Hegde N	2	U	W	P	W	W	U	U	U	U	U	U	P	ASSIGN	U	U	U	U	1
17 Hsiao A	1	ASSIGN	U	U	U	U	U	U	U	U	U	U	U	P	U	P	U	U	1
18 Kedharnath S	1	U	U	U	P	U	U	U	U	U	U	U	U	U	U	U	U	U	1
19 Kota S	2	U	ASSIGN	W	U	U	W	U	U	U	U	U	ASSIGN	U	U	U	U	U	2
20 Kulik J	1	U	U	W	U	U	W	U	U	U	U	U	ASSIGN	W	P	W	W	U	1
21 Kurra P	2	P	U	ASSIGN	P	U	U	U	U	U	U	U	U	U	U	U	U	U	1
22 Lanman K	1	U	U	U	U	ASSIGN	U	P	W	W	W	U	U	U	U	U	U	U	1
23 Levi Cc	2	P	U	U	P	U	U	W	U	U	U	U	ASSIGN	U	U	P	W	U	1
24 Merkle S	1	U	U	U	U	U	U	U	W	W	W	U	U	W	U	U	ASSIGN	P	1
25 Nidadavolu A	2	U	U	U	U	U	U	U	U	U	U	U	P	ASSIGN	U	U	P	ASSIGN	2
26 Noto S	1	P	U	U	U	U	U	U	W	ASSIGN	W	U	U	P	P	P	W	W	1
27 Nukala S	2	U	ASSIGN	W	W	W	W	W	W	W	W	ASSIGN	U	U	U	U	U	U	2
28 Pogrebitskiy D	1	W	U	U	U	U	U	U	U	U	U	U	P	U	P	W	U	U	1
29 Ritchenson S	1	U	U	U	U	U	U	U	U	U	U	U	U	ASSIGN	P	U	U	U	1
30 Rivera A	1	U	U	U	U	ASSIGN	U	P	W	U	U	U	U	U	U	U	W	U	1
31 Senthil Kumaran C	1	W	W	P	W	P	W	U	U	U	U	U	P	W	P	W	W	ASSIGN	1
32 Shah F	1	U	U	ASSIGN	P	U	U	U	W	U	U	U	W	W	U	U	U	U	1
33 Shenoy P	1	W	W	P	ASSIGN	W	P	U	U	U	U	U	P	P	W	W	U	U	1
34 Shiflett A	1	U	U	U	U	U	U	U	ASSIGN	W	W	U	U	P	P	P	U	U	1
35 Shroff R	2	U	U	U	U	U	U	W	U	U	U	U	P	W	W	W	ASSIGN	U	1
36 Sreepada S	1	U	U	U	U	U	U	U	W	W	ASSIGN	U	U	U	U	U	U	U	1
37 Stochaj K	1	W	U	U	U	U	U	U	U	U	U	U	ASSIGN	W	W	W	U	U	1
38 Sukhatankar S	1	W	P	W	U	U	U	U	U	U	U	W	ASSIGN	W	U	U	U	U	1
39 Thakkar P	2	P	U	U	U	U	U	U	U	U	U	W	W	W	W	W	ASSIGN	1	
40 Tunney B	2	U	W	P	P	ASSIGN	W	U	U	U	U	U	P	W	U	U	U	U	1
41 Waterson R	1	ASSIGN	U	U	W	W	U	W	U	U	U	U	U	W	W	W	W	W	1
42 Weigel V	1	U	U	U	U	U	U	U	U	U	U	U	ASSIGN	W	U	U	W	W	1
43 Yang R	0	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	0
Assigned:		4	4	2	3	2	4	2	2	4	3	2	3	2	2	2	4	3	
Target		4	4	2	4	2	4	2	2	4	3	2	3	2	2	2	4	3	
Gap		0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	



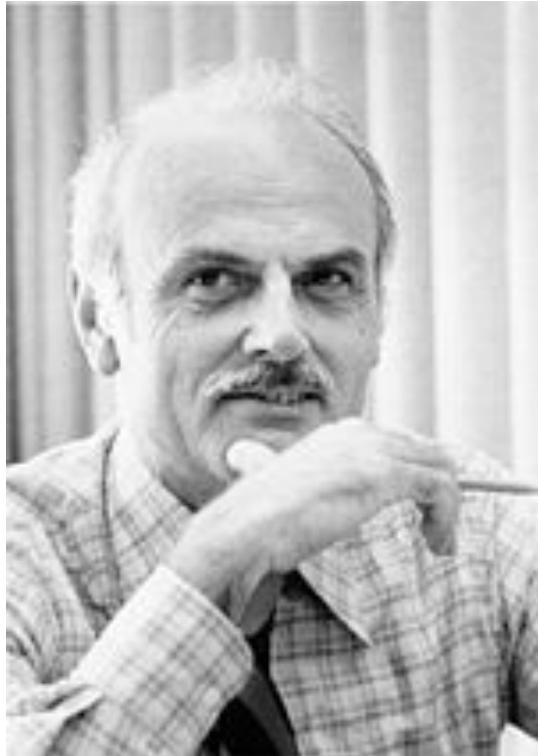
Northeastern University

Databases

Relational and Non-Relational (NoSQL)
Data Storage and Retrieval



Codd, 1970. The origins of the relational model



Edgar Frank Codd
1923-2003

Information Retrieval

A Relational Model of Data for Large Shared Data Banks

E. F. CODD

IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.



Relational Data Model

Entities

Branch

branchNo	street	city	postCode
B005	22 Deer Rd	London	SW1 4EH
B007	16 Argyll St	Aberdeen	AB2 3SU
B003	163 Main St	Glasgow	G11 9QX
B004	32 Manse Rd	Bristol	BS99 1NZ
B002	56 Clover Dr	London	NW10 6EU

Staff

staffNo	fName	lName	position	sex	DOB	salary	branchNo
SL21	John	White	Manager	M	1-Oct-45	30000	B005
SG37	Ann	Beech	Assistant	F	10-Nov-60	12000	B003
SG14	David	Ford	Supervisor	M	24-Mar-58	18000	B003
SA9	Mary	Howe	Assistant	F	19-Feb-70	9000	B007
SG5	Susan	Brand	Manager	F	3-Jun-40	24000	B003
SL41	Julie	Lee	Assistant	F	13-Jun-65	9000	B005

A column or **attribute** or **field** providing more details about the particular entity.

- Name
- Datatype
- Other attributes

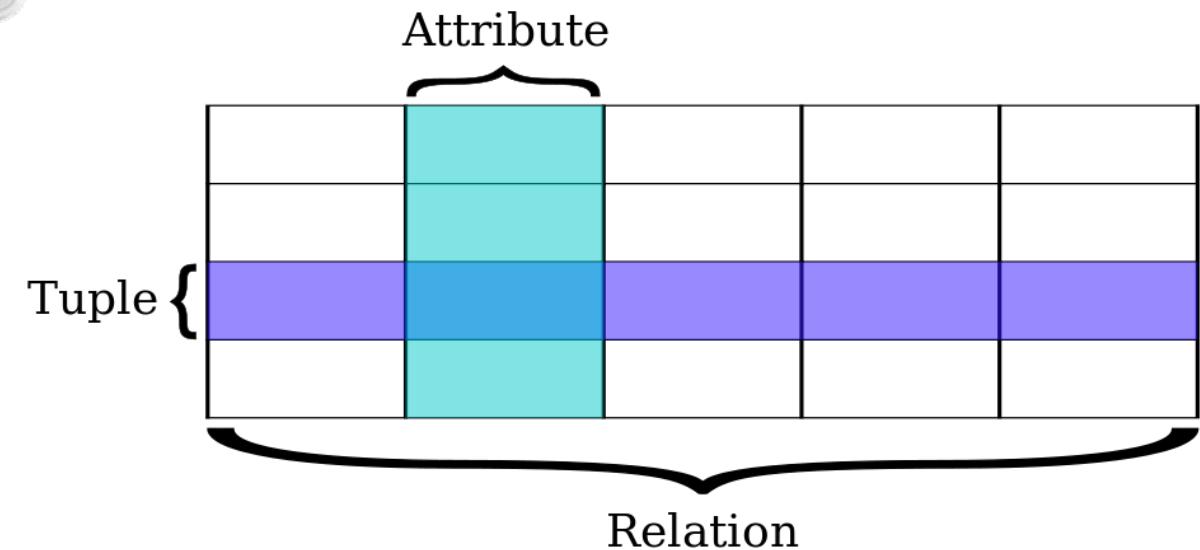
A row or **record** describing one instance of the entity

Three pillars of Relational Databases

relational model

SQL

ACID



DB-Engines Ranking

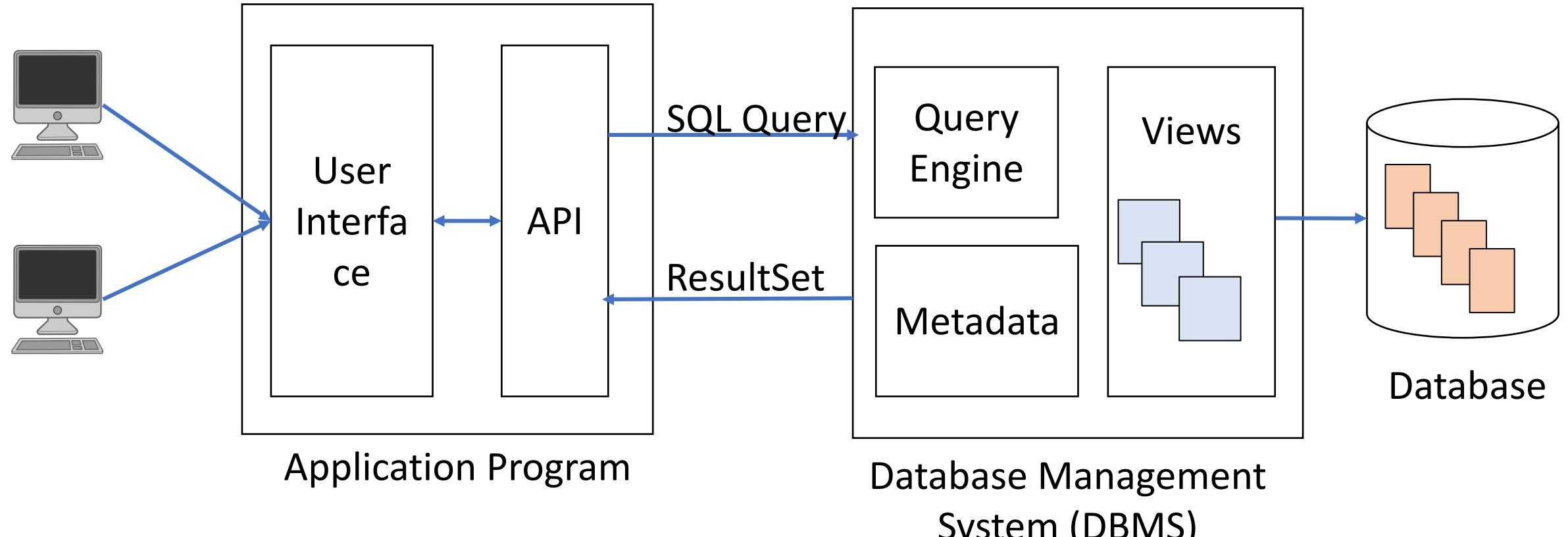
371 systems in ranking, May 2021

Rank			DBMS	Database Model	Score		
May 2021	Apr 2021	May 2020			May 2021	Apr 2021	May 2020
1.	1.	1.	Oracle	Relational, Multi-model	1269.94	-4.98	-75.50
2.	2.	2.	MySQL	Relational, Multi-model	1236.38	+15.69	-46.26
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	992.66	-15.30	-85.64
4.	4.	4.	PostgreSQL	Relational, Multi-model	559.25	+5.73	+44.45
5.	5.	5.	MongoDB	Document, Multi-model	481.01	+11.04	+42.02
6.	6.	6.	IBM Db2	Relational, Multi-model	166.66	+8.88	+4.02
7.	7.	↑ 8.	Redis	Key-value, Multi-model	162.17	+6.28	+18.69
8.	8.	↓ 7.	Elasticsearch	Search engine, Multi-model	155.35	+3.18	+6.23
9.	9.	9.	SQLite	Relational	126.69	+1.64	+3.66
10.	10.	10.	Microsoft Access	Relational	115.40	-1.33	-4.50

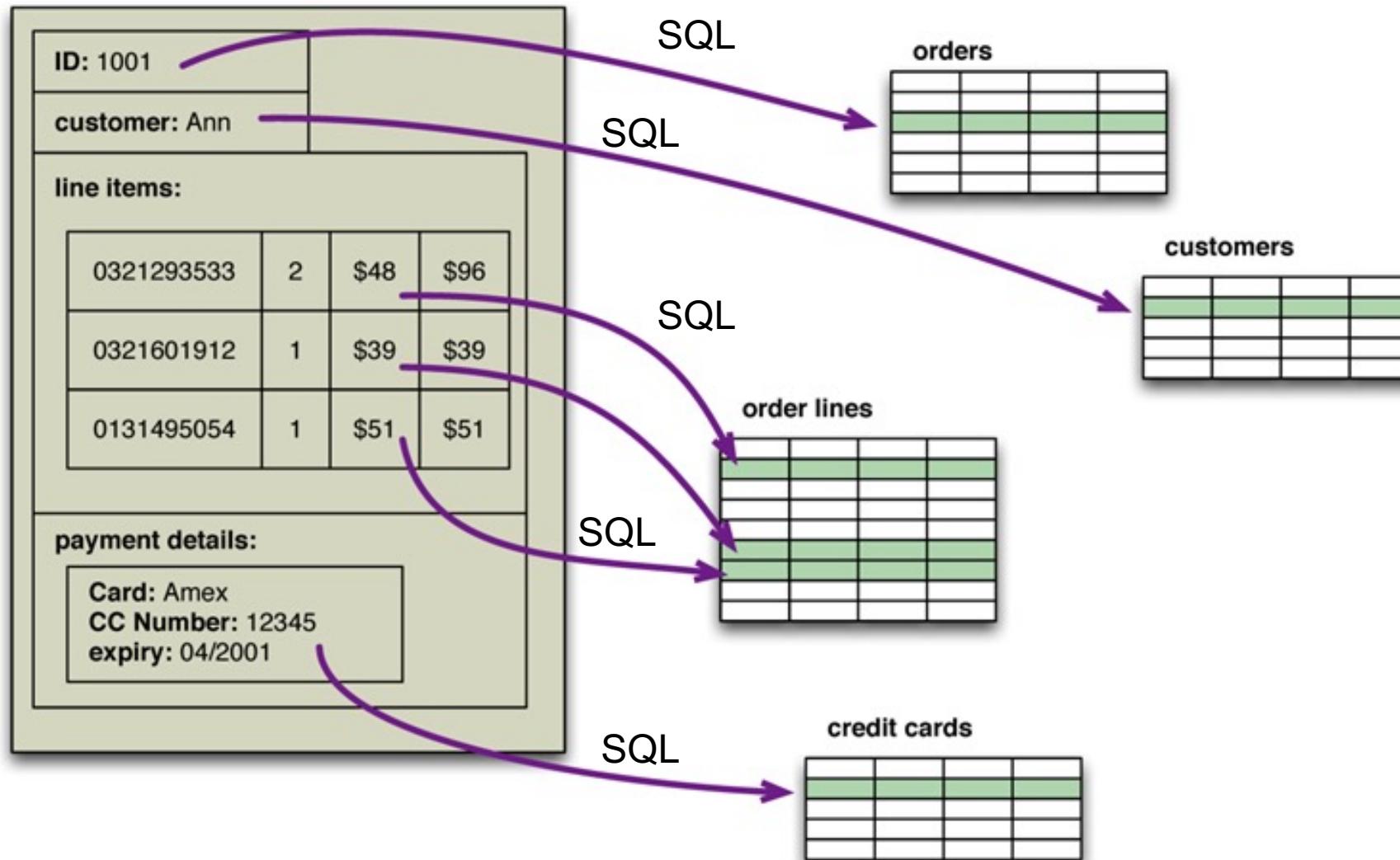
<https://db-engines.com/en/ranking>



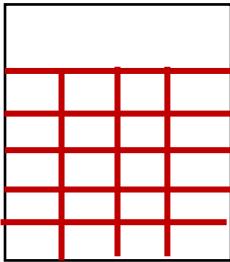
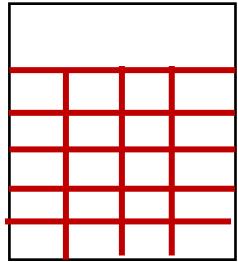
Database Application



Why NoSQL?: Impedance mismatch problem



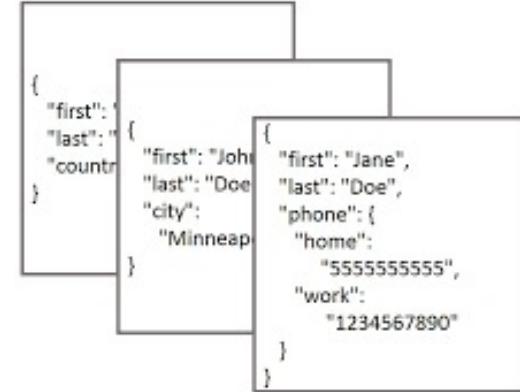
Types of models



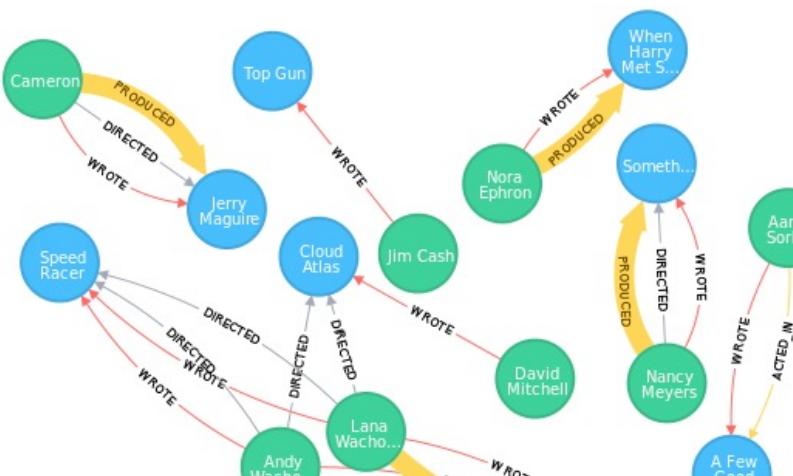
Relational

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

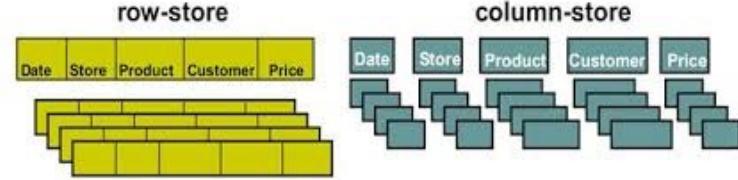
Key-Value



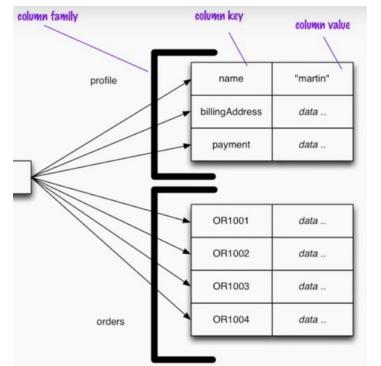
Document



Graphs



Column-oriented storage
Column-family storage

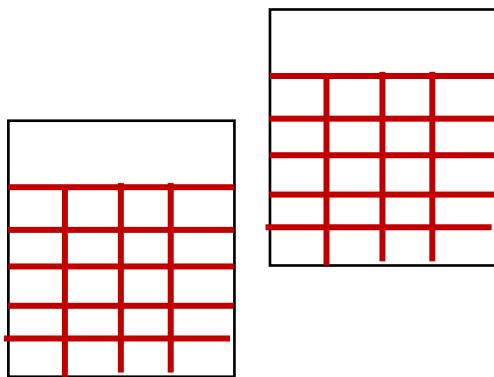


Document databases (MongoDB): a happy medium

Relational Databases

Pre-defined / rigid schema

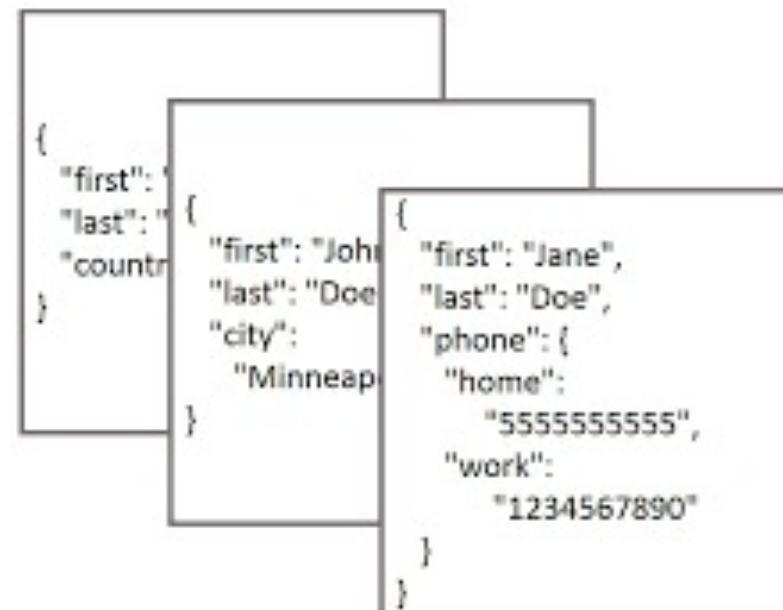
SQL: Structured Query Language



Document Databases

Data is self-describing and flexible

Query by content



Key-Value Stores

Data is simple hash tables

Query by key only (with exceptions)

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623



Document Databases

A **document database** is a non-relational database that stores data as structured documents, usually XML or JSON formats.

Designed to be *simple*, *flexible*, and *scalable*

```
{ "users": [ {  
    "firstName": "Ray",  
    "lastName": "Villalobos",  
    "joined": {  
        "month": "January",  
        "day": 12,  
        "year": 2012  
    },  
    {  
        "firstName": "John",  
        "lastName": "Jones",  
        "joined": {  
            "month": "April",  
            "day": 28,  
            "year": 2010  
        }  
    }  
]
```



Product Catalogs / e-commerce websites

The screenshot shows a product page for an Italian Shoemakers Napa Sandal. The main image is a dark brown leather sandal with a blue strap and a wooden cutout heel. Below it are smaller thumbnail images of the shoe from different angles.

Product images (points to the main product image and thumbnail images)

General Information (points to the item number and UPC information)

Localized Description (points to the localized product description and bullet points about the shoe's features)

Average Overall Rating (points to the 5-star rating icon)

Product Details (points to the detailed product description and size/color/width/quantity selection area)

List of Variants (points to the dropdown menu for selecting size, width, color, and quantity)

External Information (points to the 'Find It In Store' button and social sharing links)

Italian Shoemakers Napa Sandal

\$49.95

Compare at \$72.00

1. Select Size [size chart](#)

5.5 6 6.5 7 7.5 8 8.5 9

9.5 10 11

2. Select Width [M](#)

3. Select Color [\[Blue\]](#) [\[Black\]](#) [\[Grey\]](#)

4. Select Quantity [1](#)

ADD TO BAG

ADD TO WISH LIST

Find It In Store

Share: [f](#) [p](#) [t](#)

NEED HELP? Chat with a Shoe Lover ↗

Runs Short: Runs Long:
Runs Narrow: Runs Wide:
Not Comfy: Very Comfy:

[Read Reviews](#) [Write a Review](#) [Printable Reviews](#)

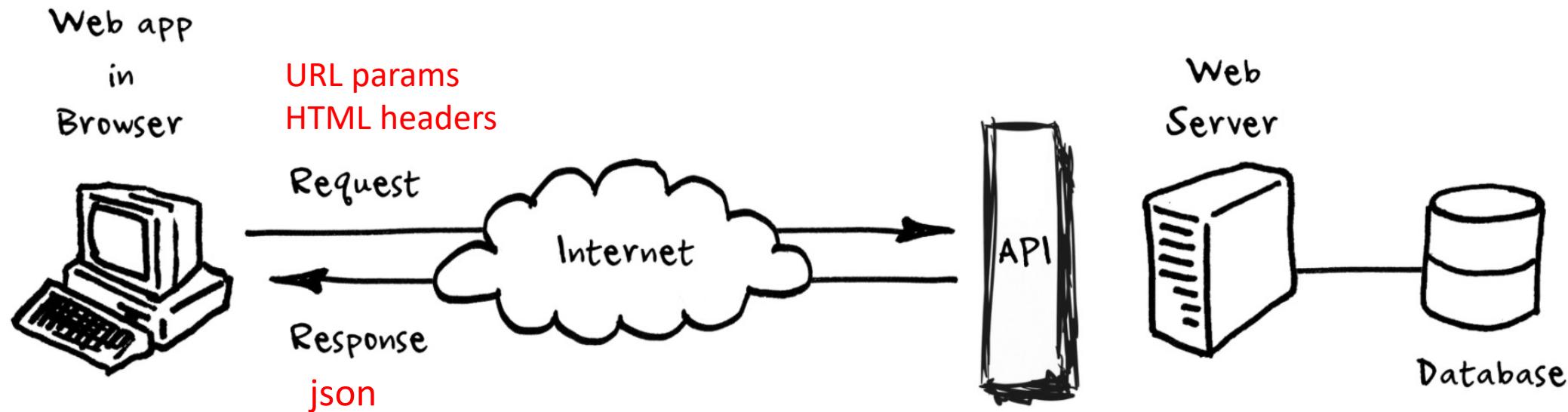
Item # 295144
UPC # 885655921272

When the weather gets hot, stay cool in the Napa platform sandal from Italian Shoemakers. This simple slide is the perfect match for your favorite sundress!

- Leather upper
- Square open toe
- ¾" platform, 3½" wood cutout heel
- Synthetic sole
- Imported
- [View more Women's Italian Shoemakers Shoes](#)

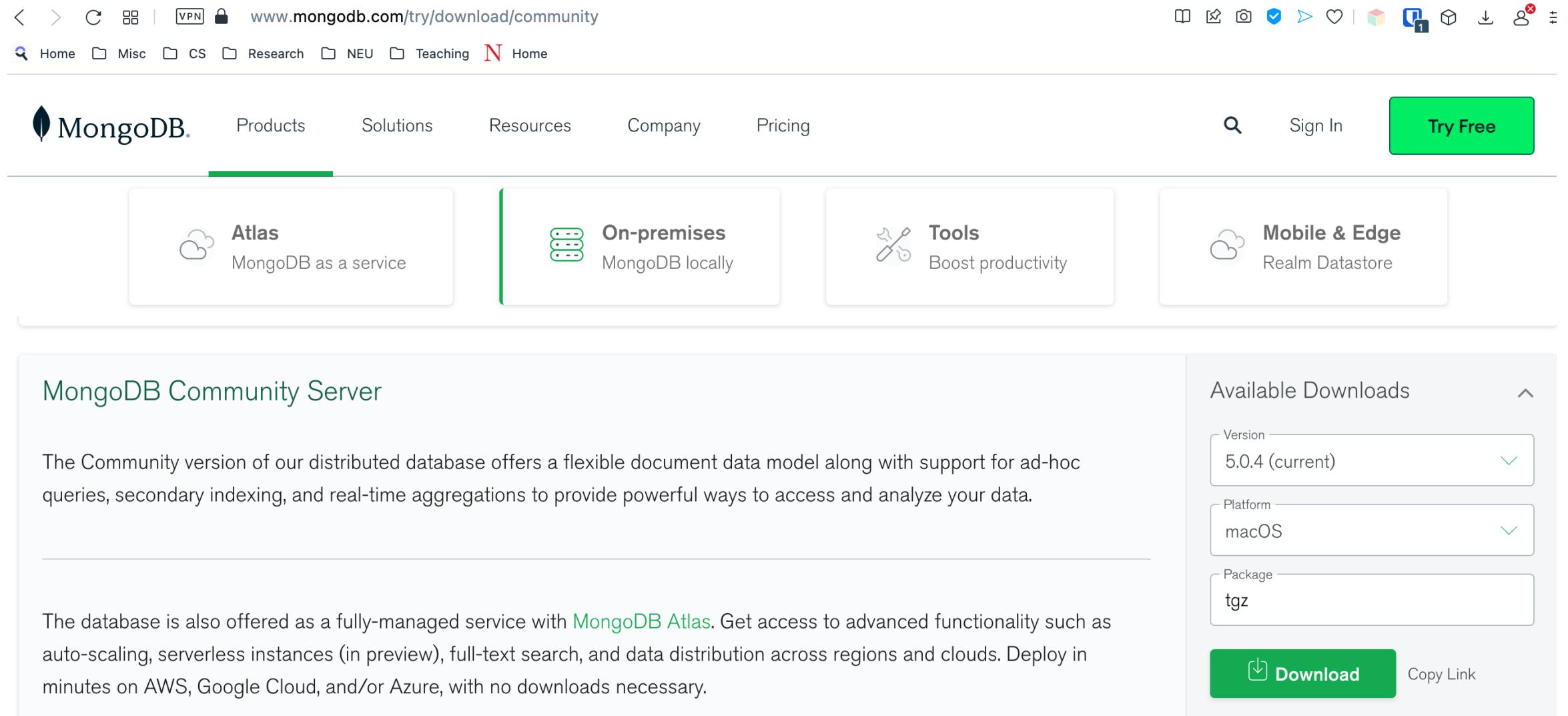
Web-Service APIs

Application Programming Interfaces (APIs) provide a general means of interacting with other programs. More specifically, many websites will provide data via a web API -- this means we can make *requests* to remote applications / websites for data. Here's what this looks like schematically.



The way that we do this in Python is via the `requests` module. The idea is that the Response (in the above depiction) is nicely formatted `json`, which feels a lot like the dictionaries you now know and love.

MongoDB Installation



The screenshot shows the MongoDB website at www.mongodb.com/try/download/community. The top navigation bar includes links for Home, Misc, CS, Research, NEU, Teaching, and a Home button. The main menu features sections for Products, Solutions, Resources, Company, Pricing, a search icon, Sign In, and a prominent green 'Try Free' button. Below this, four service options are listed: Atlas (MongoDB as a service), On-premises (MongoDB locally), Tools (Boost productivity), and Mobile & Edge (Realm Datastore). The 'On-premises' option is highlighted with a green border. The 'MongoDB Community Server' section describes the free version's features, mentioning its flexible document data model, support for ad-hoc queries, secondary indexing, and real-time aggregations. It also notes the availability of MongoDB Atlas as a managed service. To the right, a 'Available Downloads' panel allows users to select version 5.0.4 (current) for macOS in tgz package format, with a 'Download' button and a 'Copy Link' option.

Products Solutions Resources Company Pricing

Try Free

Atlas MongoDB as a service

On-premises MongoDB locally

Tools Boost productivity

Mobile & Edge Realm Datastore

MongoDB Community Server

The Community version of our distributed database offers a flexible document data model along with support for ad-hoc queries, secondary indexing, and real-time aggregations to provide powerful ways to access and analyze your data.

The database is also offered as a fully-managed service with [MongoDB Atlas](#). Get access to advanced functionality such as auto-scaling, serverless instances (in preview), full-text search, and data distribution across regions and clouds. Deploy in minutes on AWS, Google Cloud, and/or Azure, with no downloads necessary.

Available Downloads

Version: 5.0.4 (current)

Platform: macOS

Package: tgz

[Download](#) Copy Link



MongoDB Detailed Instructions

<https://docs.mongodb.com/manual/administration/install-community/>

[Install MongoDB](#) > [Install MongoDB Community Edition](#)



Install MongoDB Community Edition

These documents provide instructions to install MongoDB Community Edition.

[Install on Linux](#)

Install MongoDB Community Edition and required dependencies on Linux.

[Install on macOS](#)

Install MongoDB Community Edition on macOS systems from Homebrew packages or from MongoDB archives.

[Install on Windows](#)

Install MongoDB Community Edition on Windows systems and optionally start MongoDB as a Windows service.



MongoDB Compass: <https://www.mongodb.com/try/download/compass>

 MongoDB. Products Solutions Resources Company Pricing Sign In Try Free

 **Atlas**
MongoDB as a service

 **On-premises**
MongoDB locally

 **Tools**
Boost productivity

 **Mobile & Edge**
Realm Datastore

MongoDB Compass

Easily explore and manipulate your database with Compass, the GUI for MongoDB. Intuitive and flexible, Compass provides detailed schema visualizations, real-time performance metrics, sophisticated querying abilities, and much more.

Please note that MongoDB Compass comes in three versions: **a full version** with all features, **a read-only version** without write or delete capabilities, and **an isolated edition**, whose sole network connection is to the MongoDB instance.

For more information, see our [documentation pages](#).

 **Compass**
The full version of MongoDB Compass,
with all features and capabilities.

 **Readonly Edition**
This version is limited strictly to read
operations, with all write and delete
capabilities removed.

Available Downloads

Version: 1.29.4 (Stable) 

Platform: OS X 64-bit (10.10+) 

Package: dmg

 **Download** Copy Link

[Documentation](#)
[Archived releases](#)



Running MongoDB

Add \$MONGO_HOME to your path

```
export APPS=/Users/rachlin/apps  
export MONGO_HOME=$APPS/mongo/mongodb-4.0.7  
export PATH=$MONGO_HOME/bin:$PATH
```

Set up a directory for the database storage and run server

```
$ mkdir -p $HOME/data/mongodb  
$ mongod --dbpath $HOME/data/mongodb
```

Run the command-line client (or launch Compass)

```
$ mongo
```

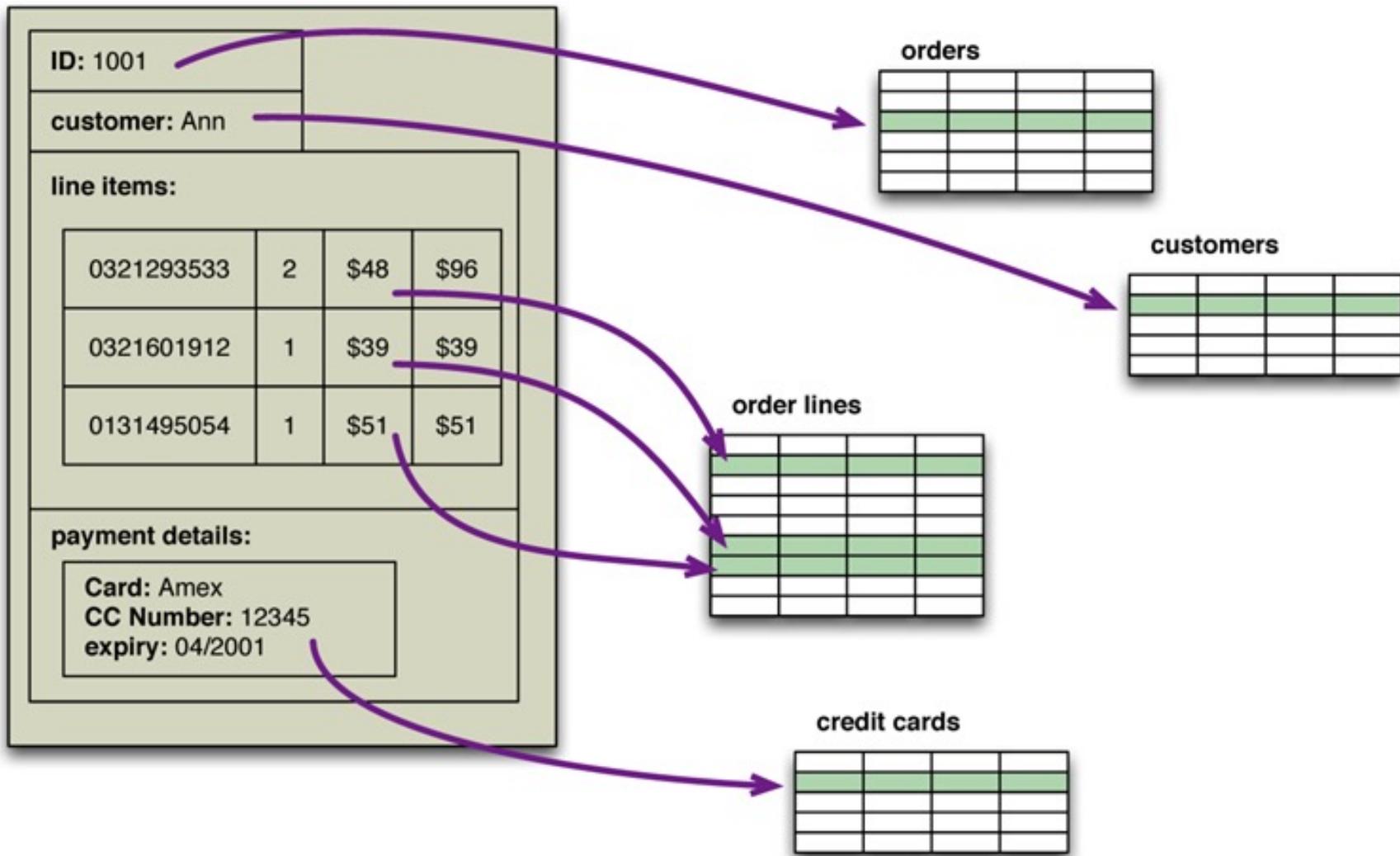


Object Persistence

Using Pickle



Impedance mismatch problem



Serialization / Deserialization

- Convert data type into a linear byte stream
- Useful for storing data or sending over network
- Python has a number of built-in modules for this
 - `marshall`
 - `json`
 - `pickle`

source: realpython.com



The marshall module (No!)

- Oldest of the three serialization modules
- Primarily reads and writes compiled bytecode from Python modules
- `.pyc` files
- Don't use - mainly used by the interpreter and can have breaking changes



JSON (human-readable serialization)

- Newest of the three serialization modules
- Produces standard JSON output: human-readable
- Works very well with other languages
- Only works with certain data types



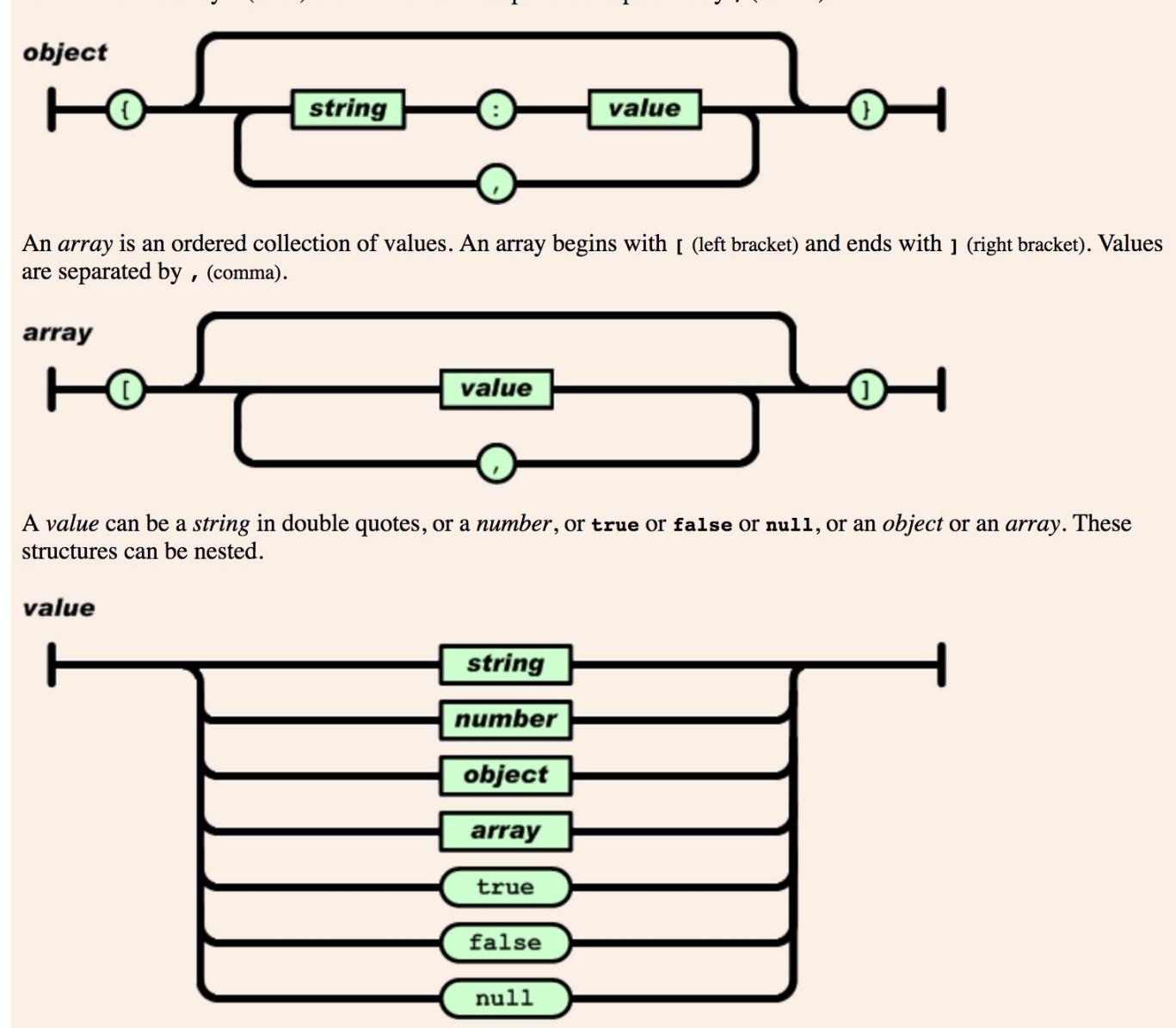
JSON

JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate.

JSON is built on two structures:

- A collection of name/value pairs. In various languages, this is realized as an *object*, record, struct, dictionary, hash table, keyed list, or associative array.
- An ordered list of values. In most languages, this is realized as an *array*, vector, list, or sequence.

These are universal data structures. Virtually all modern programming languages support them in one form or another. It makes sense that a data format that is interchangeable with programming languages also be based on these structures.



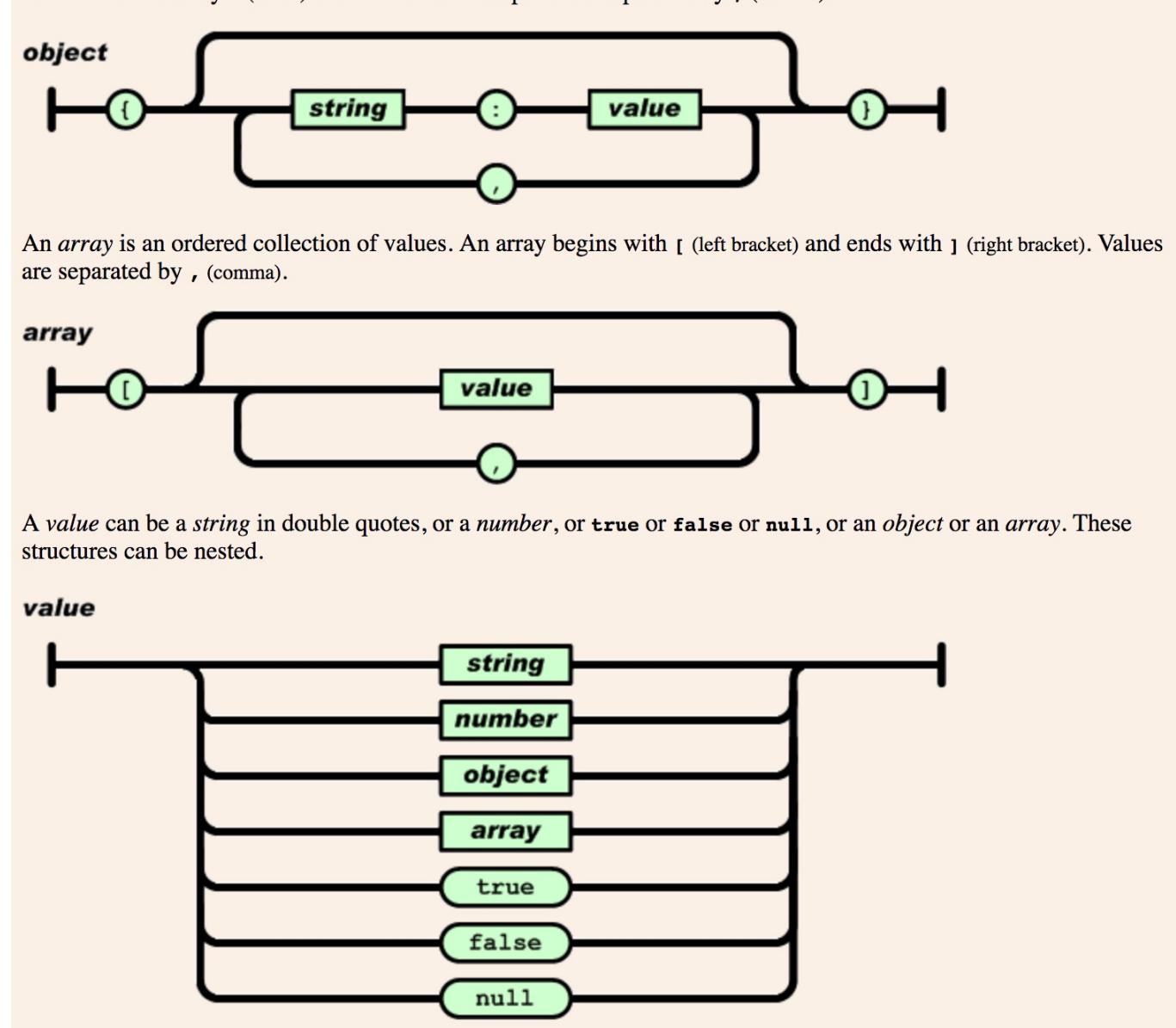
JSON

JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate.

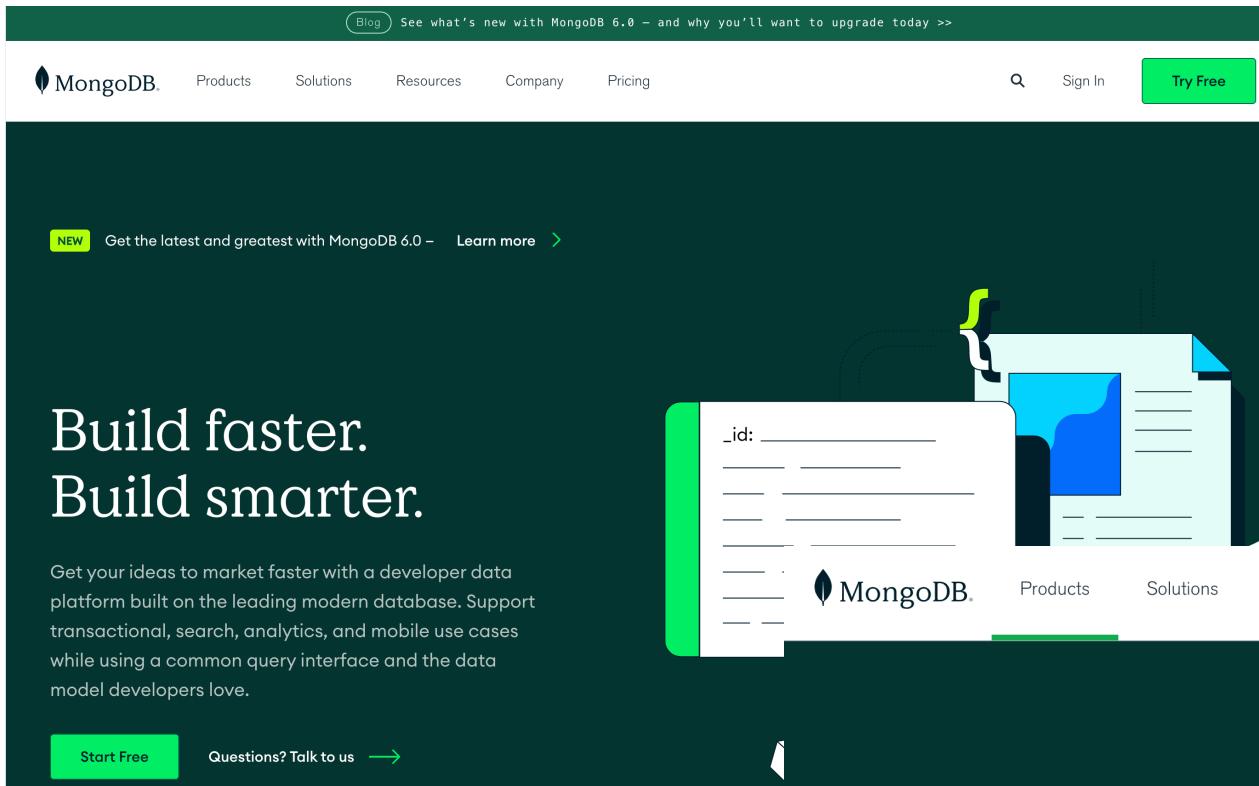
JSON is built on two structures:

- A collection of name/value pairs. In various languages, this is realized as an *object*, record, struct, dictionary, hash table, keyed list, or associative array.
- An ordered list of values. In most languages, this is realized as an *array*, vector, list, or sequence.

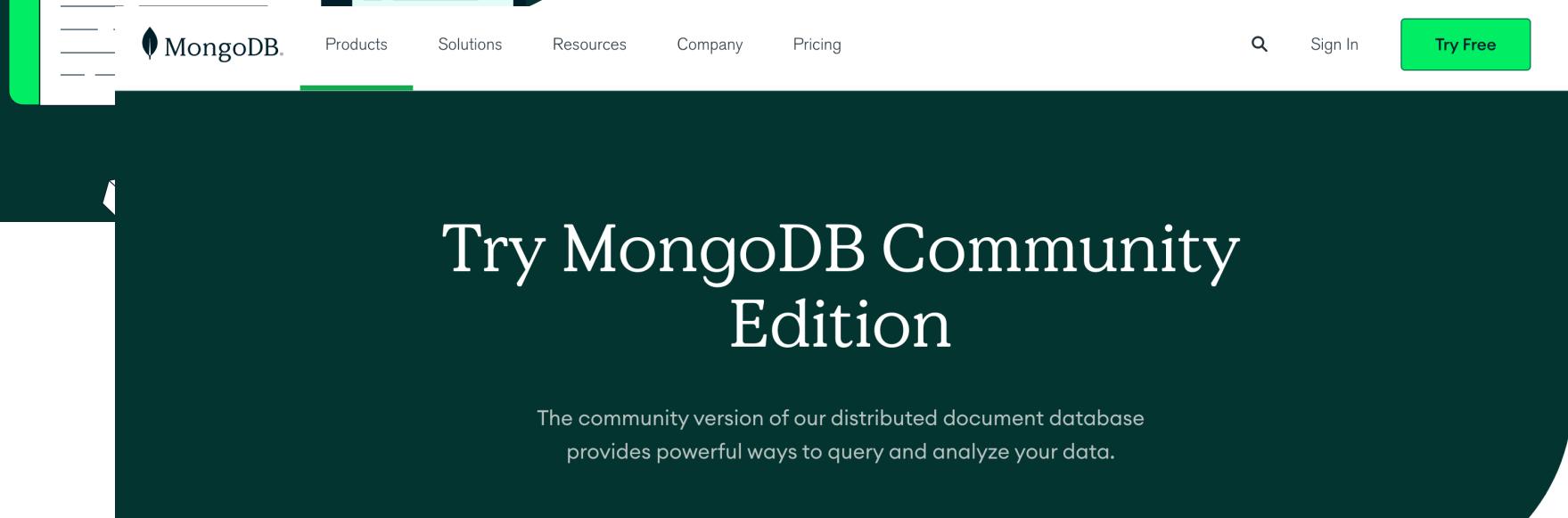
These are universal data structures. Virtually all modern programming languages support them in one form or another. It makes sense that a data format that is interchangeable with programming languages also be based on these structures.



Mongo: A Document Store that uses Binary JSON



The screenshot shows the official MongoDB website. At the top, there's a green header bar with a "Blog" button, a "See what's new with MongoDB 6.0 – and why you'll want to upgrade today >>" link, and a search bar. Below the header, the MongoDB logo is on the left, followed by navigation links for "Products", "Solutions", "Resources", "Company", and "Pricing". To the right of the navigation are a search icon, "Sign In", and a prominent green "Try Free" button. The main content area features a dark background with a large white text block that reads "Build faster. Build smarter." and "Get your ideas to market faster with a developer data platform built on the leading modern database. Support transactional, search, analytics, and mobile use cases while using a common query interface and the data model developers love.". Below this text are two calls-to-action: a green "Start Free" button and a link "Questions? Talk to us →".



A large, dark teal call-to-action banner. It features the MongoDB logo at the top left and a large white text block in the center that reads "Try MongoDB Community Edition". Below this text is a smaller white paragraph: "The community version of our distributed document database provides powerful ways to query and analyze your data." At the very bottom left of the page, there's a small circular logo for "Northeastern University".



Northeastern University

BSON (Binary JSON)



BSON [*bee · sahn*], short for Binary [JSON](#), is a binary-encoded serialization of JSON-like documents. Like JSON, BSON supports the embedding of documents and arrays within other documents and arrays. BSON also contains extensions that allow representation of data types that are not part of the JSON spec. For example, BSON has a Date type and a BinData type.

BSON can be compared to binary interchange formats, like [Protocol Buffers](#). BSON is more "schema-less" than Protocol Buffers, which can give it an advantage in flexibility but also a slight disadvantage in space efficiency (BSON has overhead for field names within the serialized data).

BSON was designed to have the following three characteristics:

1. **Lightweight**

Keeping spatial overhead to a minimum is important for any data representation format, especially when used over the network.

2. **Traversable**

BSON is designed to be traversed easily. This is a vital property in its role as the primary data representation for [MongoDB](#).

3. **Efficient**

Encoding data to BSON and decoding from BSON can be performed very quickly in most languages due to the use of C data types.



BSON Specification 1.1

Specification Version 1.1

BSON is a binary format in which zero or more ordered key/value pairs are stored as a single entity. We call this entity a *document*.

The following grammar specifies version 1.1 of the BSON standard. We've written the grammar using a pseudo-[BNF](#) syntax. Valid BSON data is represented by the `document` non-terminal.

Basic Types

The following basic types are used as terminals in the rest of the grammar. Each type must be serialized in little-endian format.

<code>byte</code>	1 byte (8-bits)
<code>int32</code>	4 bytes (32-bit signed integer, two's complement)
<code>int64</code>	8 bytes (64-bit signed integer, two's complement)
<code>uint64</code>	8 bytes (64-bit unsigned integer)
<code>double</code>	8 bytes (64-bit IEEE 754-2008 binary floating point)
<code>decimal128</code>	16 bytes (128-bit IEEE 754-2008 decimal floating point)

Non-terminals

The following specifies the rest of the BSON grammar. Note that quoted strings represent terminals, and should be interpreted with C semantics (e.g. "\x01" represents the byte 0000 0001). Also note that we use the * operator as shorthand for repetition (e.g. (""\x01"*2) is "\x01\x01"). When used as a unary operator, * means that the repetition can occur 0 or more times.

<code>document ::= int32 e_list "\x00"</code>	BSON Document. <code>int32</code> is the total number of bytes comprising the document.
<code>e_list ::= element e_list</code>	



JSON → BSON

The following are some example documents (in JavaScript / Python style syntax) and their corresponding BSON representations.

{ "hello": "world" }	→	\x16\x00\x00\x00 \x02 hello\x00 \x06\x00\x00\x00world\x00 \x00	// total document size // 0x02 = type String // field name // field value // 0x00 = type EOO ('end of object')
{ "BSON": ["awesome", 5.05, 1986] }	→	\x31\x00\x00\x00 \x04BSON\x00 \x26\x00\x00\x00 \x02\x30\x00\x08\x00\x00\x00awesome\x00 \x01\x31\x00\x33\x33\x33\x33\x33\x33\x14\x40 \x10\x32\x00\xc2\x07\x00\x00 \x00 \x00	



Embedded vs. References Data Models

```
{  
    "_id" : ObjectId("59769440c5160d8a42a02358"),  
    "firstName" : "John",  
    "lastName" : "Cent",  
    "age" : 45,  
    "email" : "john.cent@abc.com",  
    "address" : [  
        {  
            "city" : "NY",  
            "state" : "NY",  
            "country" : "USA"  
        }  
    ]  
}
```

```
{  
    "_id" : ObjectId("59769578c5160d8a42a0235a"),  
    "firstName" : "John",  
    "lastName" : "Cent",  
    "age" : 45,  
    "email" : "john.cent@abc.com",  
    "address" : [  
        ObjectId("597694f5c5160d8a42a02359")  
    ]  
}
```



```
{  
    "_id" : ObjectId("597694f5c5160d8a42a02359"),  
    "city" : "NY",  
    "state" : "NY",  
    "country" : "USA"  
}
```



Product Catalogs / e-commerce websites

The screenshot shows a product page for an Italian Shoemakers Napa Sandal. The main image is a dark brown leather sandal with a blue strap and a wooden cutout heel. Below it are smaller thumbnail images of the shoe from different angles.

Product images (points to the main product image and thumbnail images)

General Information (points to the item number and UPC information)

Localized Description (points to the product details and localized description)

Average Overall Rating (points to the star rating and review links)

External Information (points to the 'Find It In Store' and 'Share' buttons)

List of Variants (points to the size, width, color, and quantity selection fields)

Product Details
Item # 295144
UPC # 885655921272

When the weather gets hot, stay cool in the Napa platform sandal from Italian Shoemakers. This simple slide is the perfect match for your favorite sundress!

- Leather upper
- Square open toe
- ¾" platform, ¾" wood cutout heel
- Synthetic sole
- Imported
- [View more Women's Italian Shoemakers Shoes](#)

Average Overall Rating
★★★★★

Runs Short: Runs Long:
Runs Narrow: Runs Wide:
Not Comfy: Very Comfy:

[Read Reviews](#) [Write a Review](#) [Printable Reviews](#)

Italian Shoemakers Napa Sandal
\$49.95
Compare at \$72.00

1. Select Size [size chart](#)
 5.5 6 6.5 7 7.5 8 8.5 9
 9.5 10 11

2. Select Width
 M

3. Select Color
 Blue Black Grey

4. Select Quantity
 1

ADD TO BAG **ADD TO WISH LIST**

Find It In Store Share: [f](#) [p](#) [t](#)

NEED HELP?
[Chat with a Shoe Lover](#)



Flexible product descriptions

➤ db.item.findOne()

```
{ _id: "301671", // main item id
  department: "Shoes",
  category: "Shoes/Women/Pumps",
  brand: "Guess",
  thumbnail: "http://cdn.../pump.jpg",
  image: "http://cdn.../pump1.jpg", // larger version of thumbnail
  title: "Evening Platform Pumps",
  description: "Those evening platform pumps put the perfect
finishing touches on your most glamourous night-on-the-town
outfit",
  shortDescription: "Evening Platform Pumps",
  style: "Designer",
  type: "Platform",
  rating: 4.5, // user rating
  lastUpdated: Date("2014/04/01"), // last update time
  ...
}
```

Product Search

- **Get item by id**

```
db.item.findOne( { _id: "301671" } )
```

- **Get item from Product Ids**

```
db.item.findOne( { _id: { $in: ["301671", "301672"] } } )
```

- **Get items by department**

```
db.item.find({ department: "Shoes" })
```

- **Get items by category prefix**

```
db.item.find( { category: /^Shoes\|Women/ } )
```

- **Indices**

productId, department, category, lastUpdated



pickle (binary serialization)

- Serializes in binary format, not human-readable
- Works out of the box with many Python data types - including custom
- Very fast



pickle protocols

- Six different protocols for `pickle`
- Tied to Python interpreter version (2.3, 3.0, 3.4, 3.8)
- Newer protocols need newer interpreter
- Highest level available will be used, except `protocol=parameter`

